A generalized approach to generate synthetic short-to-medium range hydro-meteorological forecasts

Zachary Paul Brodeur[1] and Scott Steinschneider[1]

[1]Cornell University

November 24, 2022

Abstract

Forecast informed reservoir operations holds great promise as a soft pathway to improve water resources system performance. Methods for generating synthetic forecasts of hydro-meteorological variables are crucial for robust validation of this approach, as numerical weather prediction hindcasts are only available for a relatively short period (10-40 years) that is insufficient for assessing risk related to forecast-informed operations during extreme events. We develop a generalized error model for synthetic forecast generation that is applicable to a range of forecasted variables used in water resources management. The approach samples from the distribution of forecast errors over the available hindcast period and adds them to long records of observed data to generate synthetic forecasts. The approach utilizes the flexible Skew Generalized Error Distribution (SGED) to model marginal distributions of forecast errors that can exhibit heteroskedastic, auto-correlated, and non-Gaussian behavior. An empirical copula is used to capture covariance between variables and forecast lead times and across space. We demonstrate the method for medium-range forecasts across Northern California in two case studies for 1) streamflow and 2) temperature and precipitation, which are based on hindcasts from the NOAA/NWS Hydrologic Ensemble Forecast System (HEFS) and the NCEP GEFS/R V2 climate model, respectively. The case studies highlight the flexibility of the model and its ability to emulate space-time structures in forecasts at scales critical for flood management. The proposed method is generalizable to other locations and computationally efficient, enabling fast generation of long synthetic forecast ensembles that are appropriate for water resources risk analysis.

A generalized approach to generate synthetic short-to-medium range hydro-meteorological forecasts

Zachary P. Brodeur[1], Scott Steinschneider[2]
Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY

1. Graduate Research Assistant, 111 Wing Drive, Riley-Robb Hall, Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY, 14853. Email: zpb4@cornell.edu, Phone: 607-255-2155 (Corresponding Author).

2. Assistant Professor, 111 Wing Drive, Riley-Robb Hall, Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY, 14853. Email: ss3378@cornell.edu, Phone: 607-255-2155.

Abstract

Forecast informed reservoir operations holds great promise as a soft pathway to improve water resources system performance. Methods for generating synthetic forecasts of hydro-meteorological variables are crucial for robust validation of this approach, as numerical weather prediction hindcasts are only available for a relatively short period (10-40 years) that is insufficient for assessing risk related to forecast-informed operations during extreme events. We develop a generalized error model for synthetic forecast generation that is applicable to a range of forecasted variables used in water resources management. The approach samples from the distribution of forecast errors over the available hindcast period and adds them to long records of observed data to generate synthetic forecasts. The approach utilizes the flexible Skew Generalized Error Distribution (SGED) to model marginal distributions of forecast errors that can exhibit heteroskedastic, auto-correlated, and non-Gaussian behavior. An empirical copula is used to capture covariance between variables and forecast lead times and across space. We demonstrate the method for medium-range forecasts across Northern California in two case studies for 1) streamflow and 2) temperature and precipitation, which are based on hindcasts from the NOAA/NWS Hydrologic Ensemble Forecast System (HEFS) and the NCEP GEFS/R V2 climate model, respectively. The case studies highlight the flexibility of the model and its ability to emulate space-time structures in forecasts at scales critical for flood management. The proposed method is generalizable to other locations and computationally efficient, enabling fast generation of long synthetic forecast ensembles that are appropriate for water resources risk analysis.

1. Introduction

Forecast informed reservoir operations have tremendous potential to enhance the efficient use of water resources, especially where systems are operated near their limits or where inflows are highly variable across timescales (Faber and Stedinger, 2001; Anghilieri et al., 2016; Turner et al., 2017; Jasperse et al., 2017; Nayak et al., 2018, Guiliani et al., 2019; Delaney et al., 2020). Forecast informed policies are often developed using climate and hydrologic hindcasts, or retrospective forecasts based on models that are initialized using initial conditions that were present over a historical period. Hindcasts are split into calibration and testing periods, such that policies are designed and/or optimized for the calibration period and then tested out of sample using testing period data. A major challenge with this approach is that it is limited to the available hindcast period, which is often constrained to the era when satellite data can be used for climate model initialization (i.e., from 1979 onward; Hartmann, 2016). This limitation requires that a fairly short time period (at most ~40 years) be parsed into even smaller periods to enable calibration and testing of policies, creating the potential for overfitting and poor out-of-sample performance (Nayak et al. 2018; Brodeur et al. 2020; Herman et al., 2020).

Synthetic forecasts offer a solution to overcome this challenge. Synthetic forecasts are generated by adding random error to observational records, such that the resulting series is statistically indistinguishable to forecasts developed using a physically-based model. Many water resources projects in the United States have instrumental records of streamflow and climate that extend back to the early 20th century (Loucks & Van-Beek, 2017). Synthetic forecasts based on these extended observational records, which often contain multiple floods and droughts, provide a rich source of information for calibrating and testing forecast informed reservoir operations.

Forecast-informed policy design can benefit from synthetic forecasts at all timescales (Denaro et al., 2017). Seasonal forecasting is often used to inform water supply based decisions (Anghilieri et al., 2016; Turner et al., 2017; Guiliani et al., 2019; Yuan et al., 2015), whereas shorter range forecasts are often confined to hazard management (Valeriano et al., 2010; You & Cai, 2008). However, uncertainty in seasonal forecasts surpasses that of short-to-medium range forecasts (Giuliani et al., 2019). In addition, regions that receive a significant portion of their annual water supply from a small number of events (e.g., California; Hanak et al., 2011; Dettinger et al., 2016) benefit more from short time scale forecasts associated with these events (Jasperse et al., 2017; Nayak et al., 2018; Raso et al., 2014). Thus, shorter lead forecasts can have more value for decision-making in many regions and further increase the likelihood that water managers would incorporate them into operations (Rayner et al., 2005). Accordingly, the present study concentrates on short to medium range (0-14 days) synthetic forecast generation.

The two primary methods for synthetic streamflow forecasting are 1) the 'direct statistical error model' approach or 2) a 'conceptual hydrologic model' approach (Lamontagne & Stedinger, 2018). The direct approach derives synthetic forecasts of streamflow based on a statistical model of the errors between streamflow forecasts and observations (Lettenmaier et al., 1984). In contrast, the conceptual approach uses hindcasts of meteorological variables (generally temperature and precipitation) to drive hydrologic forecast model outputs (Alemu et al., 2011). Where long observed streamflow records exist, the direct approach is straightforward to apply and captures both hydrologic and meteorological sources of forecast uncertainty. In many regions, extensive streamflow records are not available (Teegavarapu et al., 2019), making the conceptual approach an attractive alternative. Conceptual methods can be constraining if they are reliant on available meteorological hindcast data (~1979 to present), but synthetic meteorological forecasts can overcome this issue. Synthetic meteorological forecasts are common for energy system applications dependent on forecasts of wind and solar power generation (Sun et al., 2020; Pelland et al., 2013; Olauson et al., 2016; Hodge et al., 2012; Demello et al., 2011; Barth et al., 2011), but are relatively rare in water resources management applications (Nayak et al., 2018).

This study forwards a novel method for synthetic forecast generation that can be applied to either streamflow or meteorological data, supporting both the direct and conceptual approach to synthetic streamflow forecasting. The proposed methodology addresses two challenges in synthetic forecast development that are currently unresolved. First, hydro-meteorological forecast errors often exhibit highly non-normal distributions that can be challenging to model. We forward the generalized error distribution (GED) to address this challenge. The GED distribution has a long history in the statistical literature (Subbotin, 1923) and has been referenced in alternate forms as the 'exponential power (EP)' distribution (Box and Tiao, 1992), the 'generalized Laplace' (Ernst, 1998), and the 'generalized normal' distribution (Nadarajah, 2005). A uniting feature of all variants is that they can fit empirical distributions with varying degrees of kurtosis (Cerqueti et al., 2019). This flexibility allows modeling of fat-tailed distributions that can assign higher probabilities to large forecast errors. Non-linear responses such as rain-on-snow events (Guan et al., 2016), localized orographic enhancement of precipitation (Hecht & Cordeira, 2017; Holton & Hakim, 2013), positive feedback mechanisms in strong frontal systems (Eiras-Barca et al., 2018), or timing errors (forecast 'misses') yield a propensity towards this type of error structure in short-range hydro-meteorological forecasts. Methods to add skew to the distribution (Fernandez & Steel, 1998; Buckle, 1995**;** Botazzi & Secchi, 2011) further increase the flexibility of the model to account for asymmetries in probability mass around the mode. Use of the GED is widespread in the econometrics literature (e.g. Cerqueti et al., 2019; So et al., 2008, Nelson, 1991), where skewed and fat-tailed errors are common, but its use in the hydrologic literature is less established (with some exceptions, e.g., Schoups & Vrugt, 2010).

Another challenge is that synthetic forecasts of hydro-meteorological variables must preserve correlations across space and time and realistic uncertainty bounds at various lead times (Wilks, 2011; Demargne et al., 2009). While spatio-temporal consistency is important in synthetic forecasts at any timescale, it is particularly important in short to medium range forecasts (1-14 days) that must capture transient storms as they move across the landscape (Hartmann, 2016; Wilks, 2011). In addition, forecasts errors can also be correlated with the observed data, and these correlations should be replicated in synthetic forecast development (Lamontagne & Stedinger, 2018). Copulas have emerged in hydrology as an effective tool to capture dependence across variables and spatio-temporal domains (Chen & Guo, 2019; Teegavarapu et al., 2019 and references therein). Copulas have been used primarily to model dependencies in observed data (e.g. multi-site flood frequency, drought analyses, etc.; Chen & Guo, 2019), but they have also been used for synthetic flood forecasting (e.g., the Martingale Model of Forecast Error; Zhao et al., 2013; Heath & Jackson, 1994). We were unable to find examples of copula-based synthetic meteorological forecasts, but they have been used for a variety of climate forecast related purposes including bias correction (Piani & Haerter, 2012) and ensemble post-processing (Wilks, 2015).

This work develops an adaptable synthetic forecast generation methodology that can a) model both hydrological and meteorological forecast errors exhibiting auto-correlation, heteroscedasticity, and a variety of distributional forms, b) link those errors across space, time and/or variables with empirical copulas, and c) preserve critical relationships between the observed data and forecast errors. This approach mirrors multivariate Generalized Autoregressive Conditional Heteroskedastic (GARCH) models in the econometrics literature (Wei, 2019; Rao & Vinod, 2019) and employs the Skew Generalized Error Distribution (SGED) (Wurtz et al., 2020) as the underlying model for marginal errors. We demonstrate this approach in two separate applications for hydrological and meteorological synthetic forecast generation. In the first, we generate synthetic streamflow forecasts for Folsom Reservoir, CA, emulating those of the Hydrologic Ensemble Forecast System (HEFS) currently in operational use in NOAA/NWS River Forecast Centers (RFC) (Demargne et al., 2014). Here we focus on model performance in the temporal domain across 1-10 day forecast lead times. In the second application, we synthetically generate forecasts of temperature and precipitation based on the NCEP GEFS/R numerical weather prediction model (Hamill et al., 2013) at 5 lead times and across 30 grid cells that span Northern California. This case study highlights the ability of the approach to capture key features of conditional dependence between forecasted variables across space and time. These two applications demonstrate the ability of the generalized approach to support the direct statistical model or conceptual hydrologic model approach for synthetic streamflow generation (Lamontagne & Stedinger, 2018).

2. Data

All hydro-meteorological data were collected from a region within northern California (36.5^o – 42.5^oN and 119.5^o – 124.5^oW; Figure 1). Within this region, our first synthetic forecast application focuses on hydrologic forecasts in the American River watershed (inset of Figure 1). We obtained hindcast daily streamflow data directly upstream of the Folsom reservoir from the NOAA/NWS California/Nevada River Forecast Center (CA/NV RFC, 2020) for the period of January 1 1985 to September 15 2010 and observed streamflow data from the California Department of Water Resources Data Exchange Center (CDEC) for the period of October 1 1948 to September 15 2010 (CA/DWR, 2020). As observed data, we use full natural flow (FNF) into Folsom reservoir, which is an estimated time series of natural streamflow that has been adjusted from the gauged record to remove the impacts of upstream regulation and diversions (Zimmerman et al., 2018). This process is imperfect and negative flows are produced across approximately 8.5% of the record, which we corrected to zero flow in this study.

The hindcasted streamflow data are medium-range ensemble mean output from the HEFS. These forecasts are driven by 6-hourly meteorological forcing from the NCEP Global Ensemble Reforecast System (GEFS) version 2 ensemble mean hindcast of precipitation (PRECIP), maximum temperature (TMAX), and minimum temperature (TMIN). Streamflow hindcasts are provided at an hourly timescale and initialized at 12:00 GMT daily. The HEFS model includes a Meteorological Ensemble Forecast Processor (MEFP) that converts the raw meteorological forecast data into a 61-member bias-corrected ensemble of mean areal temperature/precipitation (MAT/MAP) that is converted to streamflow through the Hydrologic Processor, incorporating observed and forecast information from hydrologic, snowmelt, and reservoir models among others (Demargne et al., 2014). The hourly HEFS model output is aggregated to a daily scale between the time period from 08:00 AM – 08:00 PM GMT to match the observed FNF data, which is recorded at 00:00 local time. There were some missing data in the HEFS output, most notably in the period from September 16 – 30 for all years. We estimated these missing values using linear interpolation, since they often occur during times of low flow and little natural variability or for a small number of individual days scattered throughout the record.

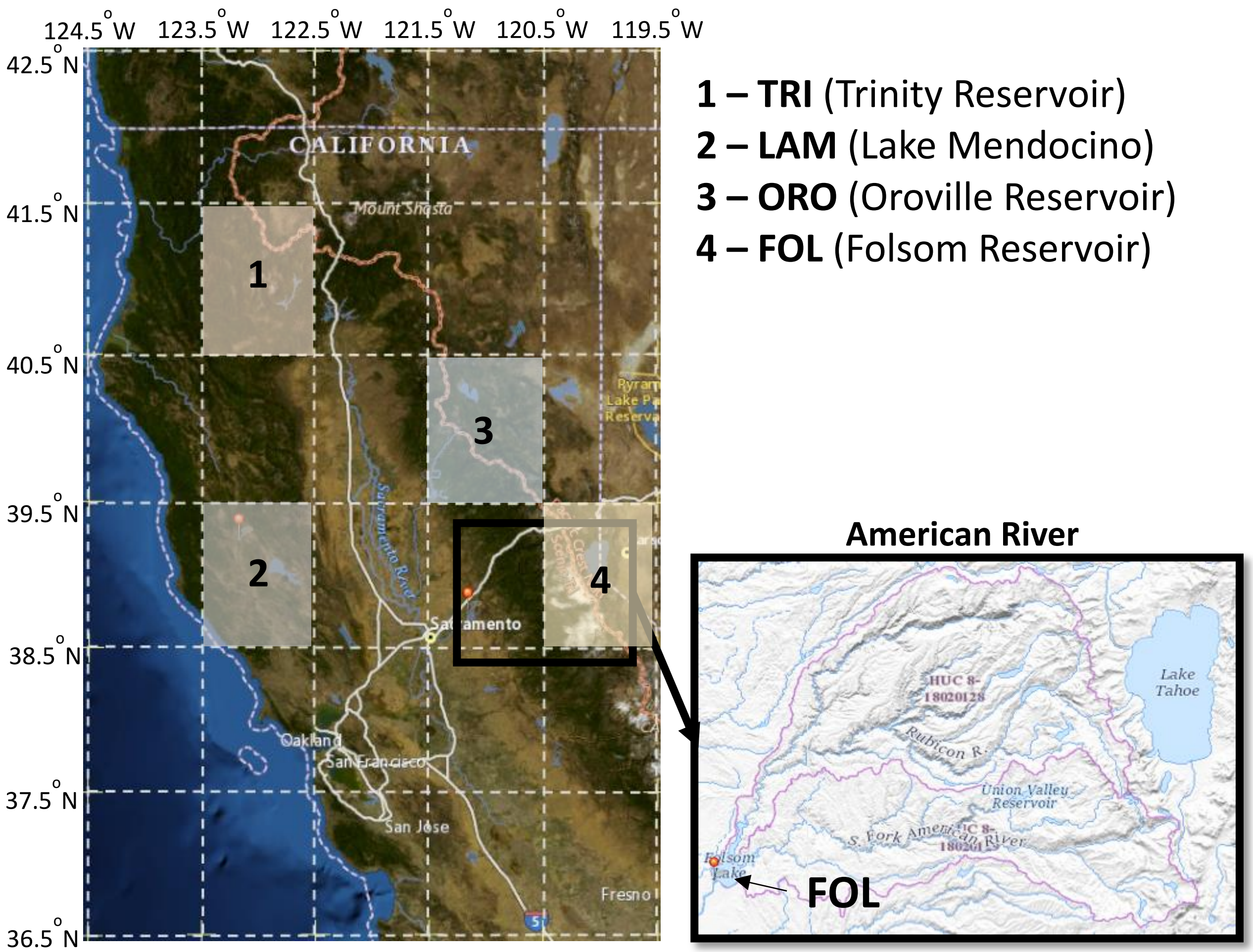

Figure 1: Geographical area of study (northern California). White dashed lines indicate the 30 grid cells used in the meteorological analysis, where the four highlighted grid cells overlay substantial portions of the watersheds explored in the meteorological case study (TRI, LAM, ORO, and FOL). The region outlined in black delimits the Folsom Reservoir watershed used in the streamflow case study, where the pink outline in the inset shows the HUC-8 sub-basin boundaries of the north and south forks of the American River (USGS, 2020).

Our second synthetic forecast application focuses on meteorological forecasts across 30 grid cells within Northern California (see Figure 1), including four grid cells that overlay the watersheds for Trinity Reservoir (TRI – 40.5^o – 41.5^oN, 122.5^o – 123.5^oN), Lake Mendocino (LAM – 38.5^o – 39.5^oN, 122.5^o – 123.5^oN), Oroville Reservoir (ORO – 39.5^o – 40.5^oN, 120.5^o – 121.5^oN), and Folsom Reservoir (FOL – 38.5^o – 39.5^oN, 119.5^o – 120.5^oN). These four locations, which we will focus on in the results, span the eastern and western slopes of the coastal ranges (TRI and LAM, respectively) and the middle and high elevations of the western slope of the Sierra Nevada range (ORO and FOL, respectively).

We obtained observed data for PRECIP, TMAX, and TMIN from the NOAA-CIRES-DOE 20th Century Version 3 (20CRV3) historical reanalysis dataset between October 1 1948 to December 31 2015 (NOAA PSL, 2020). We use reanalysis meteorology, instead of gauge-based meteorology, for its parity with the NCEP GEFS/R version 2 reforecast model (described below). The reanalysis data (hereafter referred to as observational data) are catalogued at the same spatio-temporal scale as the reforecasts (1^o x 1^o, 6-hourly) and are also produced using the NCEP GFS as the underlying model, albeit a somewhat newer version (Slivinski et al., 2019). This ensures that underlying physical processes are emulated consistently between the observational and reforecast datasets. We also note that the HEFS model applies a bias correction to forecasted meteorology (via the MEFP), so synthetic forecasts of meteorology based on reanalysis observations would be bias corrected using gauged-based meteorological data before being used to develop HEFS streamflow forecasts (Demargne et al., 2014). We obtained hindcast meteorological data from the NCEP GEFS/R V2 data repository with the same variables and time-scales as the observational data, but starting at December 1 1984 (first available hindcast date). These data come from a single ‘frozen’ version of the GEFS reforecast model across an 11-member ensemble, and we used the ensemble mean values for all variables (Hamill et al., 2013; NOAA/NCEP, 2013).

3. Generalized Synthetic Forecast Generator

The basic structure of the procedure is a linear additive error model where each member of the multivariate *n x K* set of observed data ($O_{t,j}$) is modeled as the sum of the corresponding forecast value ($F_{t,j}$) and an error component ($\varepsilon_{t,j}$), with the allowance that errors may be auto-correlated, heteroskedastic, and non-Gaussian:

$$O_{t,j} = F_{t,j} + \varepsilon_{t,j} \; where \; t \in (1,2,\dots,n) \; and \; j \in (1,2,\dots,K) \qquad \text{(Eq. 1)}$$

Here, $t=1,\dots,n$ denotes the date, and $j=1,\dots K$ can denote different lead times (e.g., 1-day ahead, 2-day ahead, etc.), locations (e.g. grid cells or sites at a single lead), or both (e.g., multiple forecast leads and locations).

We model and generate errors in three primary steps (see Figure 2). First, we use a vector auto-regressive (VAR) model to account for temporal auto-correlation within each of the *K* time-series (Section 3.1). We then fit a generalized likelihood (GL) model (Schoups & Vrugt, 2010) to each of the *K* residual series (ϵ_t) of the VAR model, accounting for heteroscedasticity and transforming the result to *K* series of random deviates (a_t) that we model with normalized SGED distributions (Section 3.2). Finally, we model the correlation of each of the *K* a_t series via an empirical copula and simulate new series of a_t using a K-nearest neighbor (KNN) and Schaake Shuffle approach. We generate synthetic forecasts errors using these simulated a_t series by inverting the process of Sections 3.1 and 3.2 (Section 3.3). The remainder of the methods (Section 3.4) describes aspects of the procedure that are specific to given variables analyzed in this study (streamflow versus meteorology).

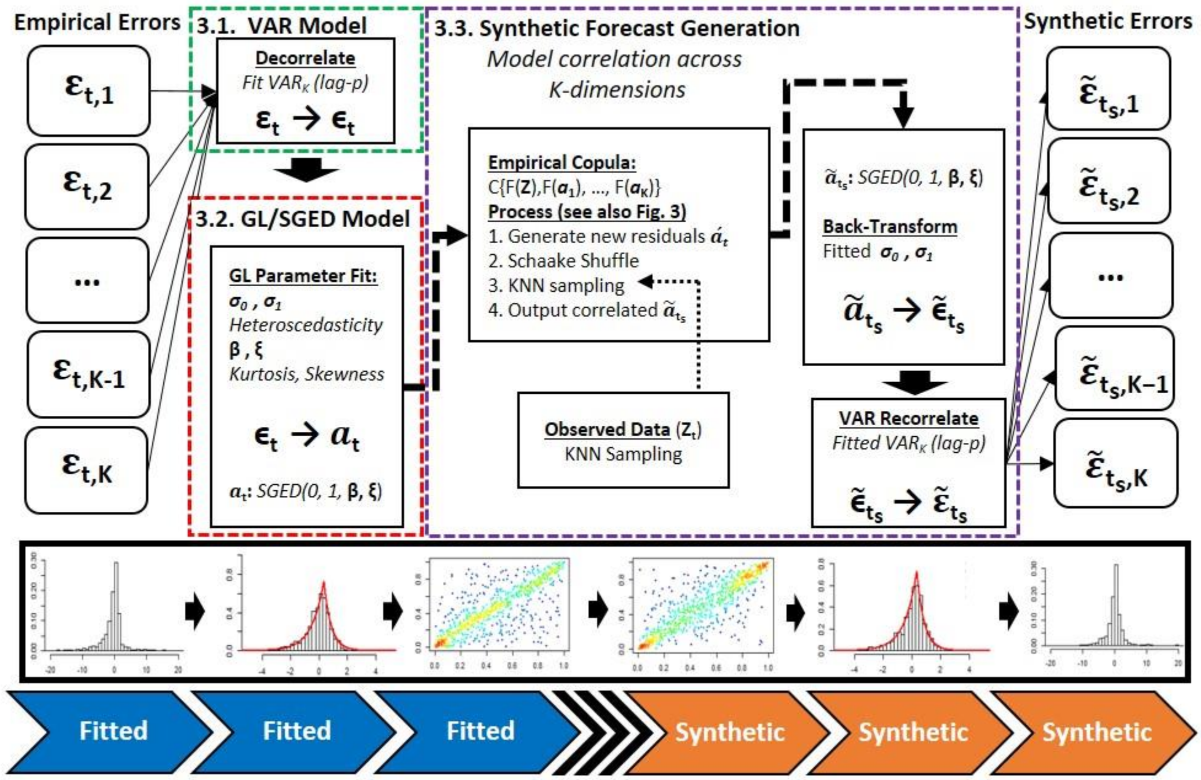

Figure 2: Conceptual model of synthetic forecast generation process, with tilde (~) notation for synthetic outputs. Section 3.1 describes the VAR model decorrelation process. Section 3.2 defines the fitting of the GL/SGED model and output of transformed residuals (a_t). Section 3.3 illustrates the synthetic generation model, with the center section showing the empirical copula approach (described in more detail in Figure 3) to model correlations across the time-series of empirical transformed residuals (a_t) and output new time-series of synthetic residuals ($\tilde{a}_{t_s}$). The right-hand section of section 3.3 then depicts the back-transformation of these residuals to raw errors ($\tilde{\varepsilon}_{t_s}$) that are used to create the synthetic forecasts. The bottom section of the figure shows graphical outputs of each step in the process (top row outlined in black) and the aligned portion of the historical period for each stage (bottom row, blue and orange chevrons).

3.1. Vector Auto-Regressive (VAR) model

We use a VAR model with *K* dimensions and lag order *p* to model auto-correlation in the forecast errors (Wilks, 2011):

$$\boldsymbol{\varepsilon}_t = \sum_{i=1}^{p}\left(\left[\,\varphi^i\right]\boldsymbol{\varepsilon}_{t-i}\right) + \boldsymbol{\epsilon}_t \qquad \text{(Eq. 2)}$$

Here, the vector of original forecast errors ($\boldsymbol{\varepsilon}_t$) at time-step *t* are equal to a linear function of lagged errors ($\boldsymbol{\varepsilon}_{t-i}$), with *K x K* matrices of coefficients ($[\varphi^i]$) out to lag order *p*. Eq. 2 reduces to a set of *K* equations that are solved via linear regression, creating both large parameter dimensions (VAR coefficients φ scale at $(Kp)^2$) and potential problems arising from high multi-collinearity (Wilks, 2011; Nicholson et al., 2020). A number of methods have been proposed to account for these issues in VAR models (Wei, 2019 and references therein); we utilize the approach of Nicholson et al., (2020) employing a group LASSO penalized model to estimate the regression coefficients while driving redundant parameters to zero. This approach selects the LASSO penalty parameter (λ) based on a rolling cross-validation and out-of-sample mean standard forecast error, helping to stabilize estimation and mitigate overfitting. We fit this model using the R-package 'BigVAR' (Nicholson et al., 2019). We used a maximal lag order (*p*) of 3, which we found was sufficient to reduce autocorrelation while maintaining model parsimony (see supporting information, Figure S1).

The residuals of the VAR model ($\epsilon_{t,j}$) are assumed to be the product of a standard random deviate ($a_{t,j}$) and an associated term to capture time-varying standard deviation ($\sigma_{t,j}$):

$$\epsilon_{t,j} = \sigma_{t,j} a_{t,j} \tag{3}$$

Models for these terms are discussed next in Section 3.2.

3.2. Generalized Likelihood and Skew Generalized Error Distribution (GL/SGED) model

We use the generalized likelihood (GL) method of Schoups and Vrugt (2010) to fit each of the *K* univariate time-series of ϵ_t to a normalized SGED distribution with corrections for heteroscedasticity. To model heteroscedasticity, the standard deviation at time *t* ($\sigma_{t,j}$) is modeled as a linear function of the observed value ($O_{t,j}$):

$$\sigma_{t,j} = \sigma_{0,j} + \sigma_{1,j} O_{t,j} \qquad \text{(Eq. 4)}$$

The estimated standard deviation $\sigma_{t,j}$ is paired with the VAR model residual $\epsilon_{t,j}$ to estimate a standardized random deviate $a_{t,j}$:

$$a_{t,j} = \epsilon_{t,j} / \sigma_{t,j} \qquad \text{(Eq. 5)}$$

These deviates $a_{t,j}$ are assumed to follow a normalized skew generalized error distribution (SGED) distribution with zero mean, unit variance, and parameterized with skew (ξ_j) and kurtosis (β_j) parameters. This is equivalent to the parameterization of the skew exponential power (SEP) distribution in Schoups and Vrugt (2010), although other parameterizations of the SGED are available (e.g., Wurtz et al., 2020). The kurtosis parameter (β_j) in this model can vary continuously between -1 and 1, where values of 1, 0, and approaching (in the limit) -1 correspond to Laplacian, Gaussian, and uniform distributions, respectively. Skewness (ξ_j) can vary continuously between 0.1 and 10 with 1 being a centered distribution and values less than 1 (greater than 1) corresponding to negative skewed (positive skewed) distributions. We use maximum likelihood estimation via numerical optimization to estimate the four parameters ($\sigma_{0,j}, \sigma_{1,j}, \beta_j, \xi_j$) simultaneously. Note that different parameters are estimated for each of the j=1,…,K series.

3.3. Synthetic Forecast Generation

To generate consistent forecast errors across space and lead time, we must preserve the correlations among the *K* series of a_t. These correlations reflect the tendency of meteorological phenomena (e.g. frontal systems, atmospheric rivers, etc.) and forecasts thereof to organize in space and time, which in turn forces space-time organization in forecast errors. In addition, it is also important to preserve the correlation between the observations and each of the $a_{t,j}$ series (Lamontagne & Stedinger, 2018). This is because actual forecasts tend to underestimate the variance of the observations, particularly if post-processed via model output statistics (MOS). However, if forecast errors are assumed to be independent of the observations, then synthetic forecasts generated by adding synthetic forecast errors to observations will have a variance greater than the actual forecasts.

The multivariate relationships between the *K* series of a_t and the observations may be difficult to model using a parametric approach (e.g., Gaussian or student-t copulas). Empirical copulas, on the other hand, preserve the observed correlation structure exactly, but random samples from an empirical copula will be limited to the range of values observed in the historic record. To address this limitation, we employ a version of the Schaake Shuffle (Clark et al., 2004), which can be interpreted as a type of empirical copula method. We use the Schaake Shuffle, coupled with a K-nearest neighbor (KNN) sampling technique, to synthesize series of $a_{t,j}$ outside of their historic range but that exhibit the same rank structure as the original data.

The steps to generate synthetic forecasts are summarized in Figure 2 and Figure 3, the latter of which provides an overview of the procedure for sampling synthetic forecast errors. We use the terminology ‘fitted period’ to refer to the time period (length *n*) in which the model parameters are estimated and ‘synthetic period’ to refer to the time period (length n_s) over which synthetic forecasts are generated. The fitted period aligns with the period containing available hindcast data, while the synthetic period aligns with the available observational period excluding the fitted period. We use the notation tilde (~) to refer to data in the synthetic period, an accent to refer to randomly generated data (e.g., $\acute{a}_{t,j}$), and the generic variable Z_t to refer to observational data used for KNN sampling. The steps to generate a synthetic forecast are as follows:

1) Rank observed $a_{t,j}$ values in each of *K* dimensions for the fitted period ($t=1, \ldots, n$)

2) Generate a new set of standardized random deviates ($\acute{a}_{t,j}$) for each dimension over the fitted period using the fitted SGED distributions

3) Rank $\acute{a}_{t,j}$ values and ‘Schaake shuffle’ to match original rank structure from step 1

4) For each time step $t_s=1, \ldots, n_s$ in the synthetic period:

    4a. Sample via KNN an observation Z_t from the fitted period based on the synthetic period observation $\tilde{Z}_{t_s}$

    4b. Populate synthetic forecast error matrix with $\acute{a}_{t,j}$ values associated with Z_t

    Result: An n_s *x K* dimension matrix composed of *K* series of (rank correlated) synthetic residuals $\tilde{a}_{t_s,j}$ over the synthetic period

5) Back-transform all $\tilde{a}_{t_s,j}$ to $\tilde{\epsilon}_{t_s,j}$ using Eqs. 4 & 5, fitted $\sigma_{0,j}$ and $\sigma_{1,j}$, and $O_{t_s,j}$

6) Convert $\tilde{\epsilon}_{t_s,j}$ to raw forecast errors $\tilde{\varepsilon}_{t_s,j}$ using Eq. 2 and fitted VAR coefficients $[\varphi^i]$

7) Convert $\tilde{\varepsilon}_{t_s,j}$ to synthetic forecasts $\tilde{F}_{t_s,j}$ using Eq. 1 and $O_{t_s,j}$

8) Repeat steps 2-7 *M* times to create *M* separate synthetic forecasts.

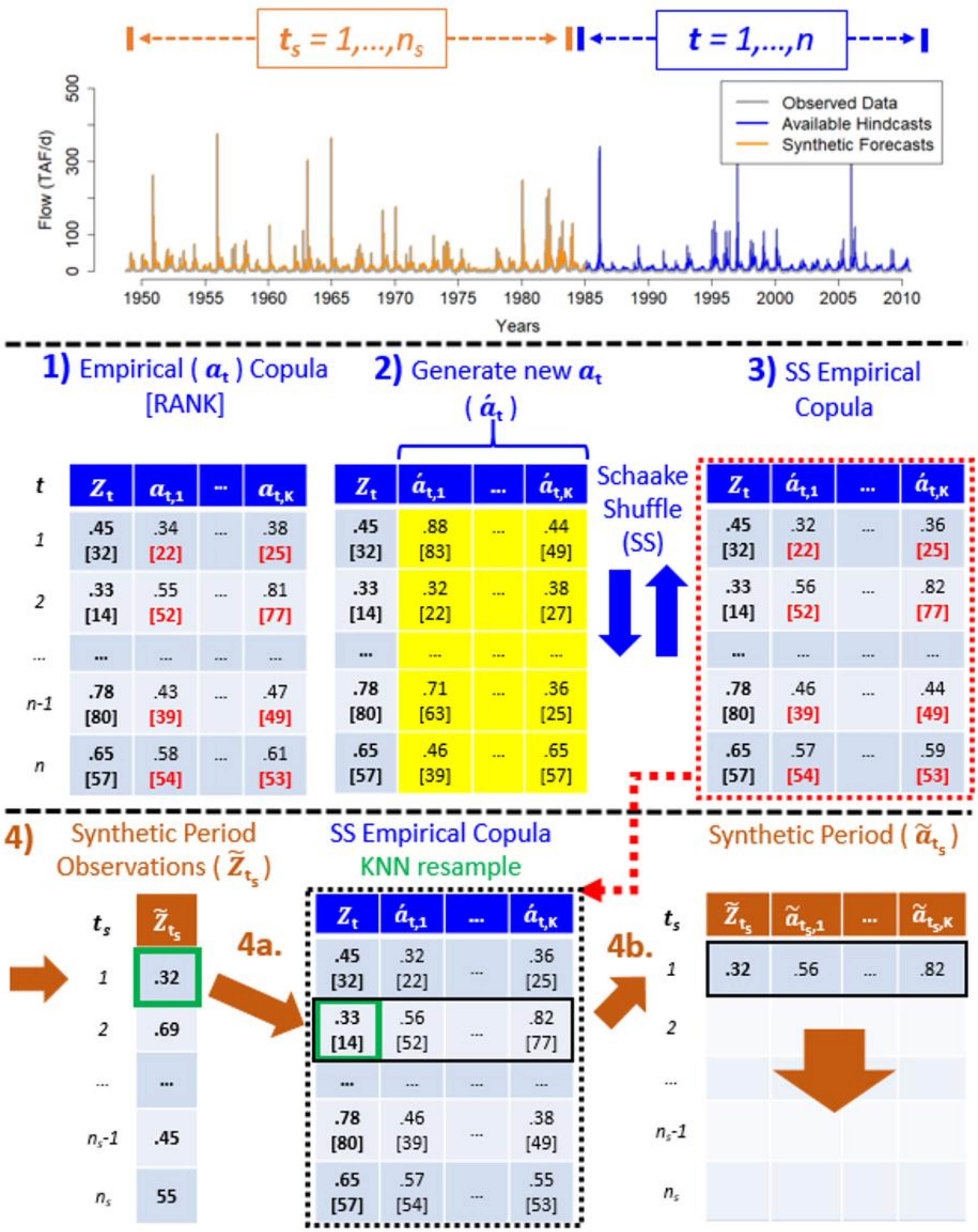

Figure 3: Graphical depiction of empirical copula/Schaake shuffle/KNN resampling procedure from section 3.3, with large bolded numbers (1,..,4b) corresponding to steps in the text. Top Row – Illustration of partitioning of observed record into 'fitted' periods (blue) and 'synthetic' periods (orange); color scheme is maintained through remainder of figure. Middle Row – Depiction of steps 1-3 of Empirical Copula/Schaake Shuffle procedure after generation of new random deviates for each of K time series ($\acute{a}_{t,j}$). Bracketed values indicate the rank of each deviate, with red bracketed text showing rank structures that are preserved in the Schaake shuffle process. Bottom Row – Illustration of step 4 for each time step t in the synthetic period. Green outline depicts KNN sampling procedure (step 4a) while population of the synthetic period standardized error matrix to length n_s is shown in step 4b.

In the procedure above, each synthetic forecast (composed of an n_s *x* K matrix of residuals over the synthetic period) is populated with randomly generated standardized residuals ($\acute{a}_{t,j}$) that are 'Schaake shuffled' to match the original rank structure over the fitted period. A row of these standardized residuals is then sampled and used for time step t_s in the synthetic period. This sampling is based on a KNN approach, whereby the synthetic period observation ($\tilde{Z}_{t_s}$) for time step t_s is used to select a value of Z_t from the fitted period (along with the randomly generated standardized residuals associated with it). The KNN procedure uses a k value of ~ $\sqrt[2]{n}$ (Lall and Sharma, 1996), a discrete kernel function (Steinschneider and Brown, 2013) to weight the k neighbors for each sample, and a Euclidean distance metric. This approach ensures that the residual values in the synthetic forecast error matrix ($\tilde{a}_{t_s,j}$) retain the correlation structure of the original empirical copula in the fitted period. The synthetic residuals $\tilde{a}_{t_s,j}$ are then converted back to raw errors ($\tilde{\varepsilon}_{t_s,j}$) and synthetic forecasts ($\tilde{F}_{t_s,j}$) by reversing the procedures in Sections 3.1 and 3.2.

It is important to note that Z_t may refer to the observations directly, or it may refer to transformations of the observed data. In particular, Z_t could be a scalar observation (e.g., observed flow), a transformed scalar observation (e.g., non-exceedance probability of observed flow), or a vector of transformed observations (e.g., principal components of precipitation occurrence). This is discussed in Section 3.4. However, when calculating the heteroscedastic components in Eq. 4 or manipulating Eq. 1, the direct observational data $O_{t,j}$ is used.

3.4. Application to Hydrology and Meteorology

The general methodology for synthetic forecasts above is applied in two case studies: 1) synthetic streamflow forecasts for Folsom Reservoir, CA that emulate the RFC HEFS forecasting system; and 2) synthetic temperature and precipitation forecasts across Northern California. Certain details of the generalized approach differ across these two applications.

For streamflow forecasts, we fit all model parameters separately by month to capture seasonal behavior in forecast residuals. The observed data used in the KNN sampling (Z_t and $\tilde{Z}_{t_s}$) is set equal to empirical non-exceedance probabilities of the observed streamflow (i.e., FNF) and calculated across the entire observation record (fitted and synthetic periods). In the KNN resampling, nearest neighbors are ambiguous when observed FNF values are zero, which occurs often in certain seasons. KNN samples in these instances are chosen randomly from all samples associated with zero observed flow values.

In the meteorological case, initial experiments (not shown) suggested that it was sufficient to fit parameters separately for the cold season (Oct.-Mar.) and warm season (Apr.-Sep.) to account for seasonality in residual behavior. We use observed precipitation occurrence for the given day (*t*) and day prior (*t-1*) as the basis for selecting nearest neighbors. For this study, we consider 30 grid cells across the Northern California region, and create a sampling matrix of dimension 60 (30 [day *t*] observations + 30 [day *t-1*] observations). We reduce this matrix to 10 dimensions (91.8% proportion of variance explained) using logistic principal component analysis (logistic PCA; Landgraf & Lee, 2015), which we implement with the 'logisticPCA' R-package (Landgraf, 2016). Then, for each precipitation occurrence observation from the synthetic period (i.e., $\tilde{Z}_{t_s}$, a vector of 10 PC values), the KNN algorithm is used to select a precipitation occurrence observation from the fitted period (Z_t, also a vector of 10 PC values), along with the associated standardized residuals ($\acute{a}_{t,j}$). As in the streamflow example, precipitation occurrence observations when there is no precipitation in any grid cell for day *t* or *t-1* are randomly sampled from all days in the fitted period where the same condition is observed. Also importantly, we impose the occurrence-based structure from each KNN sample on our synthetic forecast. That is, whatever grid cells had forecasted non-zero precipitation from the resampled day in the fitted period (Z_t) are also assumed to have forecasted non-zero precipitation for the associated synthetic day ($\tilde{Z}_{t_s}$). Then, for those grid cells with forecasted non-zero precipitation, we develop synthetic, non-zero precipitation forecasts based on synthesized residuals and the procedures in Section 3.3. This approach enables a straightforward way to capture realistic proportions of true positives/negatives and false positives/negatives from the associated forecasts. In cases where the synthetically generated residuals produce a negative precipitation forecast, we resample until a non-negative result is produced.

Finally, we note that NCEP GEFS/R V2 temperature forecasts exhibited consistently biased behavior, particularly TMIN. To improve modeling, we subtract these biases based on a monthly mean, model the resultant unbiased forecast errors, and then add the biases back in when creating the synthetic forecasts.

4. Results

4.1. Synthetic Streamflow Forecasts

We first analyzed model performance against HEFS inflow forecasts across 1-10 day lead times for Folsom reservoir. Synthetic streamflow forecast models were developed separately by month, but in the results below we selectively highlight model behavior in representative months across the year. Figure 4 shows residual behavior and some component of model fit in January and July for 1-day and 5-day lead times. Similar results for other months are presented in supporting information Figures S3-S6. The top two rows of Figure 4 show the distribution and autocorrelation function of the raw residuals (ε_t). For both months and lead times, the raw residuals are skewed and leptokurtic. The residuals in January exhibit a larger range, less autocorrelation, and a clear left skew, indicating a tendency towards over-prediction. The residual range is consistent with greater and more frequent precipitation in January that increases the chances for large streamflow errors, while a lack of snowmelt decreases error persistence. Conversely, in July, residuals are less skewed and much more persistent, reflecting the snowmelt dominated hydrology typical of the warm season in mountainous regions.

The transformed residuals (a_t) for both months and lead-times exhibit a similar distribution as the raw residuals (Figure 4, row 3), with a fat-tailed Laplacian distribution ($\beta \approx 1$) and some amount of negative skew ($\xi < 1$). Both the heteroscedastic intercept (σ_0) and scaling coefficient (σ_1) terms are substantially higher in January than July, suggesting greater baseline variability and conditional heteroscedasticity in that month, respectively. Though not shown, autocorrelation in the a_t series has been removed via the VAR model. The most notable difference between the a_t distributions across the two months is the more prominent negative skew in the January residuals, which follows the clear negative skew in the raw residuals. In both cases, the fitted SGED pdfs appears to fit the data well, but we also confirm goodness-of-fit (GoF) visually through Q-Q plots (supporting information, Figure S2).

The final two rows of Figure 4 illustrate the complicated relationship between the observed streamflow values and the a_t series (row 4), as well as the correlation between the a_t series at different lead times (row 5). These relationships are shown in terms of the empirical non-exceedance probabilities (NEPs) for all variables. The NEP values for the a_t series at 1- and 5-day lead times are clearly and strongly correlated for both months, although the dependence is weaker in January but with more upper and lower tail dependence (row 5). In contrast, the relationships between the NEPs of observed flow and the a_t series at different leads times exhibit clustered and asymmetric behavior (row 4). For example, in January, under-predicted forecasts are common for moderately sized flow events at both lead times (i.e., a_t NEPs between 0.8 – 0.9 for observed flow NEPs between 0.4 - 0.5), while over-predicted forecasts (.e.g, a_t NEPs between 0.0 – 0.4) frequently occur at moderate-to-high flow events (observed flow NEPs between 0.5 - 0.9). In July, over-predictions for very low flows are very common (i.e., a_t NEPs between 0.0-0.5 for observed flow NEPs near zero), while higher flows in July are very frequently under-estimated. These error clusters are unique to the calibration of the HEFS model and should be captured in synthetic forecast generation, which motivates our use of an empirical copula in the generation process (see Section 3.3).

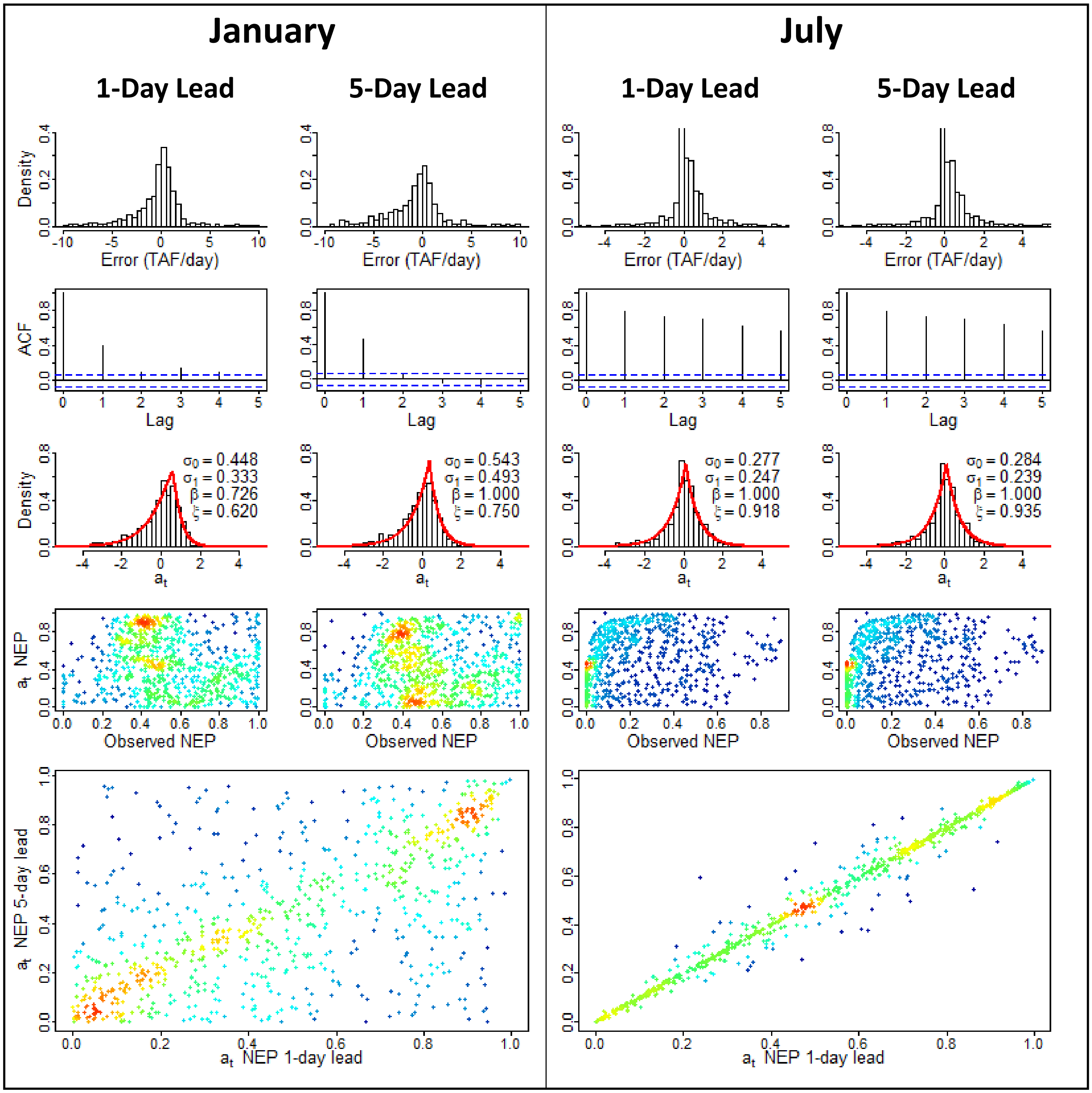

Figure 4: Graphical depiction of model fit to Folsom Reservoir hydrologic forecasts for January (July) in the left (right) halves of the overall plot. In the top 4 rows, residual fit analysis at 1-day (5-day) forecast leads are indicated in the left (right) halves of the monthly sub-sections. Top row is raw forecast residuals with the second row showing the autocorrelation function (ACF) plots. Third row shows transformed residuals (a_t) as black histogram bars with the solid red line showing the fitted SGED (0,1,β,ξ) pdf and fitted parameters indicated in black text. Fourth row shows density plots of observed non-exceedance probabilities (NEP) versus a_t NEPs, where red (blue) coloration show high (low) density. Bottom row is a density plot comparison of at NEPs between 1 and 5-day leads.

Figure 5 more clearly shows the seasonality of parameters in the GL/SGED model (and hence error structure) at lead times of 1, 3, 5, and 10 days. Seasonality in the heteroscedastic intercept (σ_0) is consistent across lead times throughout the warm season (AMJJAS), but diverges substantially in the cold season (ONDJFM), with larger values more common at longer lead times. In general, there is higher static error variance in the cold season when storms are frequent. The heteroscedastic scaling term (σ_1) is very similar across lead times, and generally is lower during months when flows are higher due to snowmelt.

The SGED kurtosis parameter (β) remains at or near 1 (Laplacian distribution) for most months and lead times, with the most noticeable exception being the 1-day lead forecasts that decrease below 1 (i.e., become more Gaussian) during the cold season. This reflects a higher probability for small to moderate forecast errors at a 1-day lead, especially during the more variable cold season months. At 1, 3, and 5 day lead times, the SGED skewness is generally negative (ξ < 1). At a 10-day lead, the skewness disappears or becomes slightly positive ($\xi \geq 1$). This shift likely reflects a tendency towards climatology in the ensemble mean forecasts at longer leads and subsequent under-predictions of large events. Finally, we note that the skewness parameter and heteroscedastic scaling term (σ_1) are negatively correlated, which likely explains some of the aberrant spikes in these parameters during certain months (August, December).

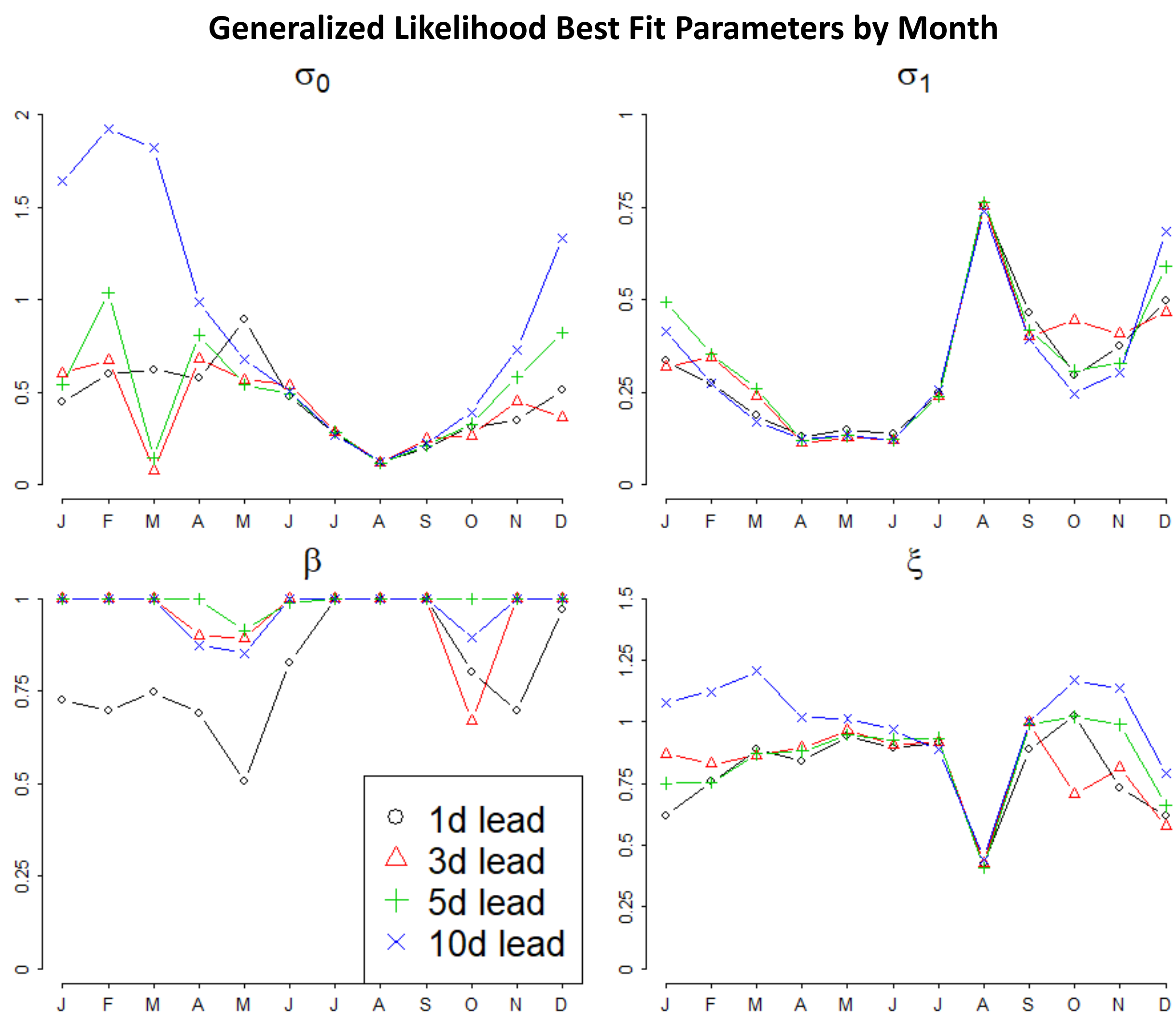

Figure 5: Fitted values (y-axis) by month (x-axis) for the four parameters of the generalized likelihood function across 1, 3, 5, and 10-day forecast leads. σ_0 and σ_1 are the intercept and slope parameters for heteroscedasticity, respectively, while β and ξ are the kurtosis and skewness parameters of the normalized SGED distribution.

To assess performance of the model for streamflow forecasts, we create 1000 synthetic forecasts over the fitted period (when hindcasts are available) and compare them to the actual HEFS forecasts. Figure 6 shows the distribution of these synthetic forecasts (expressed as 50% and 95% prediction intervals) for four months in 1986, as well as a single synthetic forecast trace. Similar results for a year in the synthetic period (1955) are shown in supporting information

(Figure S7). Several results emerge from Figure 6. First, we note that across all months and lead times, the model preserves cross-correlation in forecast error structure. For example, on January 17 1986, a large spike in the observed inflow is systematically under-predicted by the 1, 3, 5, and 10 day actual HEFS forecast (HEFS-sim). This behavior is reflected by the synthetic forecast trace (HEFS-syn) that also under-predicts across all lead times. In addition, the model captures important relationships between the observed flow and forecasts, including the general tendency for forecasts to under-predict large, infrequent events, especially at long lead-times. This is seen most clearly by the 50th percentile bounds that are depressed below the highest observed flow values, and is also confirmed through direct comparisons between observed flow and forecast residuals for both the empirical and synthetically generated data (see supporting information Figure S8).

Figure 6 also highlights how the synthetic model preserves auto-correlation in the forecasts. For instance, at a 5-day lead in April, the actual HEFS forecast (HEFS-sim) shows persistent over-predictions across the April 5-25 interval, while the synthetic forecast trace (HEFS-syn) shows a similar degree of persistence but for under-predictions. Autocorrelation appears to increase with lead time, especially in months driven by snowmelt.

The 10-day lead-time example for January 1986 displays a noteworthy limitation of the modeling structure. The empirical copula maintains the correlation between observed flows and forecast residuals, which generally lead to forecast under-predictions for large inflow events. When coupled with the autocorrelation structure imposed by the VAR model, this tends to drive even greater under-predictions in the following time steps. For very large inflow events, this can cause the synthetic forecasts to be unrealistically low, as shown by the January 10-day lead synthetic forecast trace (HEFS-syn) reaching zero flow.

For the remaining two months (July and October, row 3-4), we note many of the same characteristics as for January and April. However, the actual HEFS forecasts in these months display smoothed behavior and are increasingly detached from the variability in the observed inflow. The uncertainty bounds capture the actual forecast traces reasonably well, but the physical behavior of the synthetic forecast trace, which is tied to the observations, is substantially different than that of the smooth HEFS forecasts. We note, though, that the flow magnitudes in these months is rather low, so the practical implications of these differences in forecast behavior is likely small.

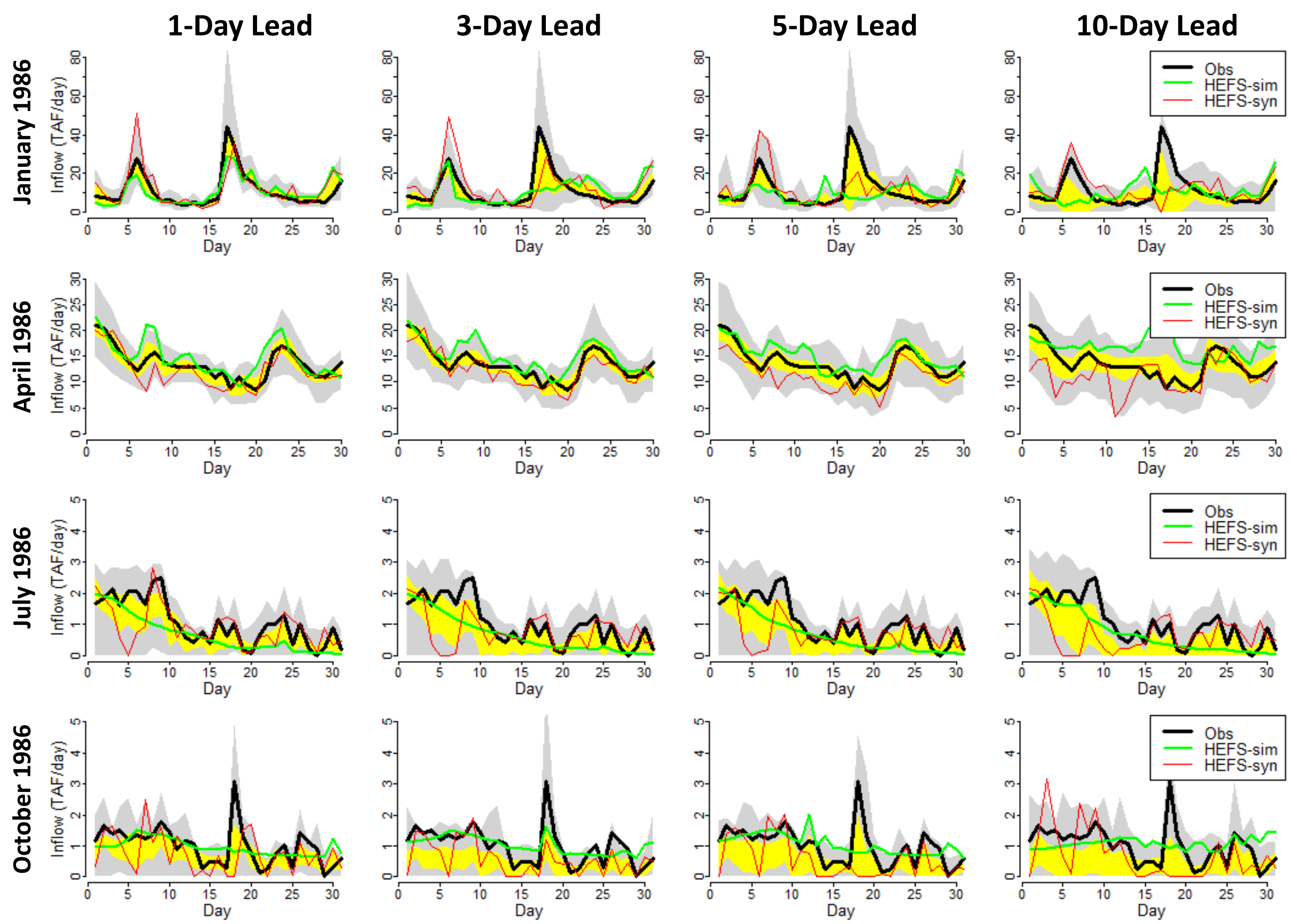

Figure 6: Folsom Reservoir (FOL) hydrographs from a selected year (1986) for four different months spaced evenly across the year showing synthetic forecast performance at 1, 3, 5, and 10 day forecast leads. Observed full natural flow is indicated by the black solid line, the HEFS ensemble mean forecast by the green solid line, and a randomly sampled synthetic forecast by the red solid line. Light grey (yellow) shading indicate the 95th (50th) percentile bounds from 1000 synthetic forecast samples.

To more systematically validate synthetic forecast performance, Figure 7 shows synthetic forecast reliability and skill. Reliability refers to the frequency that the HEFS forecasts lie within the 95% synthetic forecast bounds (which should be 95% if the synthetic forecasts were generated correctly). We assess reliability for different observed flow ranges, discretized into percentile bins (0-10th percentile, 10-20th percentile, etc.). Across lead times, January, April, and October reliability is generally near the 95% target across flow percentiles, albeit at times slightly below. The 1-day lead forecasts are slightly less reliable than the other lead times, particularly at low to moderate flows, which may be tied to the slightly more Gaussian (i.e. less fat-tailed) fits noted above for this lead-time that would lead to tighter uncertainty bounds. The July reliability diverges substantially from the other months and is usually well below 95%. Most notably, reliability is low at both the lowest and highest flows. The HEFS forecasts exhibit smoothed and sometimes biased behavior in these low-flow months (see Figure 6). This could lead to extended periods where observed flows are at or near zero and the actual HEFS forecasts are biased above the synthetic forecast uncertainty bounds, explaining the low reliability at low flows. During rare high flows in July the HEFS forecasts often significantly under-predict, and when coupled with the low error variance for this month, synthetic forecasts tend to be less able to capture these under-predicted HEFS forecasts.

Finally, we assess skill in the synthetic forecasts compared to the actual forecasts using a common mean squared error climatological skill score ($SS_{clim} = 1 - \frac{MSE}{MSE_{clim}}$; Wilks, 2011). This score captures the ability of the forecasts to outperform a climatological forecast, which in this case is a 7-day rolling average for each day of the year across the observational record. A value of 1 is a perfect forecast, a value of 0 is equivalent to climatology, and a negative value is worse than climatology. Figure 7 compares whether the synthetic forecasts match the skill of the actual HEFS forecasts for different lead times and months. January skill for the synthetic forecasts show good correspondence to actual forecasts, with the actual forecast skill remaining well within the 50th percentile bounds. April and July synthetic forecasts have too much skill compared to the actual forecasts at shorter lead times, but capture the actual forecast skill more accurately at longer leads. The overestimation of skill is particularly prevalent in July, when synthetic forecast reliability is also lowest. October underestimates skill relative to the actual forecasts, although the actual forecast skill still lies within the 95% bounds of the synthetic forecasts. Overall, the skill in the synthetic forecasts generally reflects that in the actual forecasts across months and lead times, but with some deviations that are specific to different times of year.

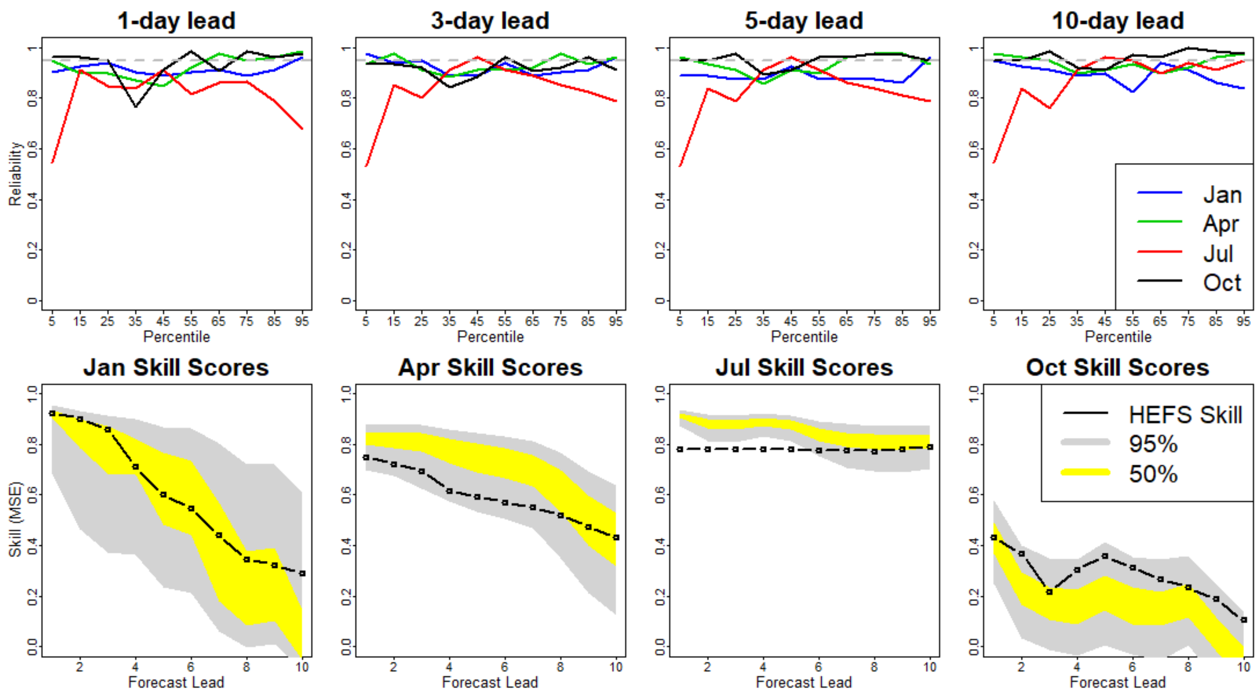

Figure 7: Top row – 95th percentile reliability plots for 1, 3, 5, and 10-day forecast leads across four selected months. The dashed grey line indicates 95% reliability while the x-axis labels show the center of each of the 10 percentile bins (i.e. '5 percentile' indicates 0 – 10 percentile values of the observed flow). Bottom row – Climatological skill score plots based on mean squared error. The black line shows the HEFS ensemble mean forecast skill across the 10 forecast leads while the gray (yellow) shading indicate 95th (50th) percentile bounds across 1000 samples.

4.2. Synthetic Meteorological Forecasts

The meteorological case study illustrates a much higher dimensional problem, as we model 3 variables (PRECIP, TMAX, TMIN) across 30 grid cells at 5 lead times, resulting in K=450 dimensions. We split the meteorological data into cold-season (ONDJFM) and warm-season subsets (AMJJAS) for model fitting, and focus our results on synthetic forecast performance during the cold season at four grid cells that overlay key watersheds (TRI, LAM, ORO, and FOL; see Figure 1). Figure 8 shows residual behavior and some components of model fit in the cold season for 2 of the 3 variables (PRECIP, TMAX) at Lake Mendocino (LAM), since TMIN behaves in a qualitatively similar manner to TMAX. Plots for additional variables, sites, and lead times are shown in supporting information Figures S9-S13. In Figure 8, PRECIP standardized forecast residuals (a_t) for the LAM site show similar distributional qualities to those for streamflow, while TMAX standardized residuals are more Gaussian (top row). The SGED model is able to capture this behavior well, and goodness-of-fit is consistent across sites and variables (see supporting information Figure S14). Both forecasts show some level of conditional heteroscedasticity ($\sigma_1 > 0$), though TMIN and TMAX are sometimes fit with no conditional heteroscedasticity at other locations/lead times. Empirical correlations between a_t NEP values at LAM and TRI (row 2) show symmetric tail-dependence typical of a t-copula (Chen & Guo, 2019) for TMAX, while PRECIP shows more pronounced lower tail dependence and other asymmetric behavior. The synthetic forecasts capture this behavior well (row 3). The final two rows of Figure 8 show raw residuals versus observed precipitation and temperature for the original hindcasts (row 4) and the synthetic forecasts (row 5). For both variables, the model accurately preserves much of the relationship between the raw residuals and the observations. We do note though that the synthetic residuals for PRECIP show slightly less variability at low observed values than the empirical residuals and vice versa for moderate to high observations.

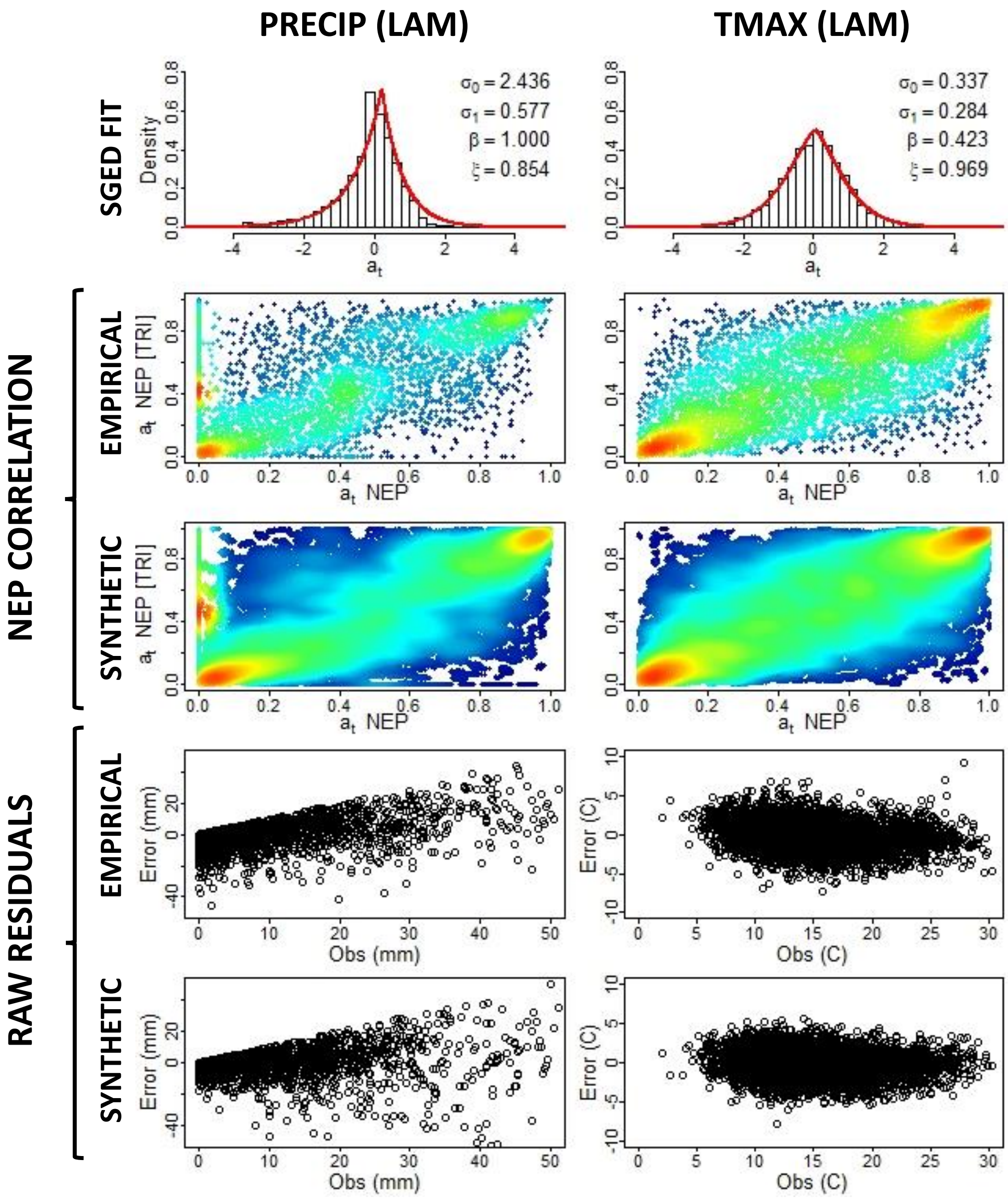

Figure 8: Model fit metrics for cold season (ONDJFM) precipitation (PRECIP – left column) and maximum temperature (TMAX – right column) at a 5-day forecast lead for LAM watershed grid cell. Top row - Fitted residuals (a_t) as in Figure 3, row 3. 2nd and 3rd row – a_t NEP correlations between LAM and TRI grid cells for empirical (synthetic) data in the 2nd (3rd) row where the 3rd row shows the density of 100 synthetic samples. 4th and 5th row – Observed values plotted against residual errors for empirical (synthetic) in the 4th (5th) row. Synthetic residual errors are shown from a single synthetic forecast sample.

Similar to our approach for streamflow, we assess synthetic meteorological forecast performance using 1000 synthetic forecast traces over the fitted period. Figure 9 shows the distribution of these synthetic forecasts for PRECIP, TMAX, and TMIN in February 1986. Similar results for another month and year from the synthetic period (December 1955) are shown in supporting information (Figure S15). In Figure 9, we note similar cross-correlated behavior in the forecasts to those of streamflow, except in this instance the correlations are spatial. For example, at all sites the sampled synthetic PRECIP forecast trace (GEFS-syn) primarily underestimates the observed event from February 13-17, and then at three sites (LAM, ORO, and FOL) it overestimates the observations between February 18-19. This shows how the synthetic forecast trace captures the synchronized error in event timing across locations.

The TMAX and TMIN GEFS forecasts exhibit less variable behavior that is well captured by the synthetic forecast model. Cross-correlations still exist (note the over-prediction of TMAX near February 14 and 22 across sites). There is some moderate negative bias in the TMIN forecasts that is especially evident at the LAM location, which the synthetic forecasts are able to capture through a simple bias correction (see Section 3.4).

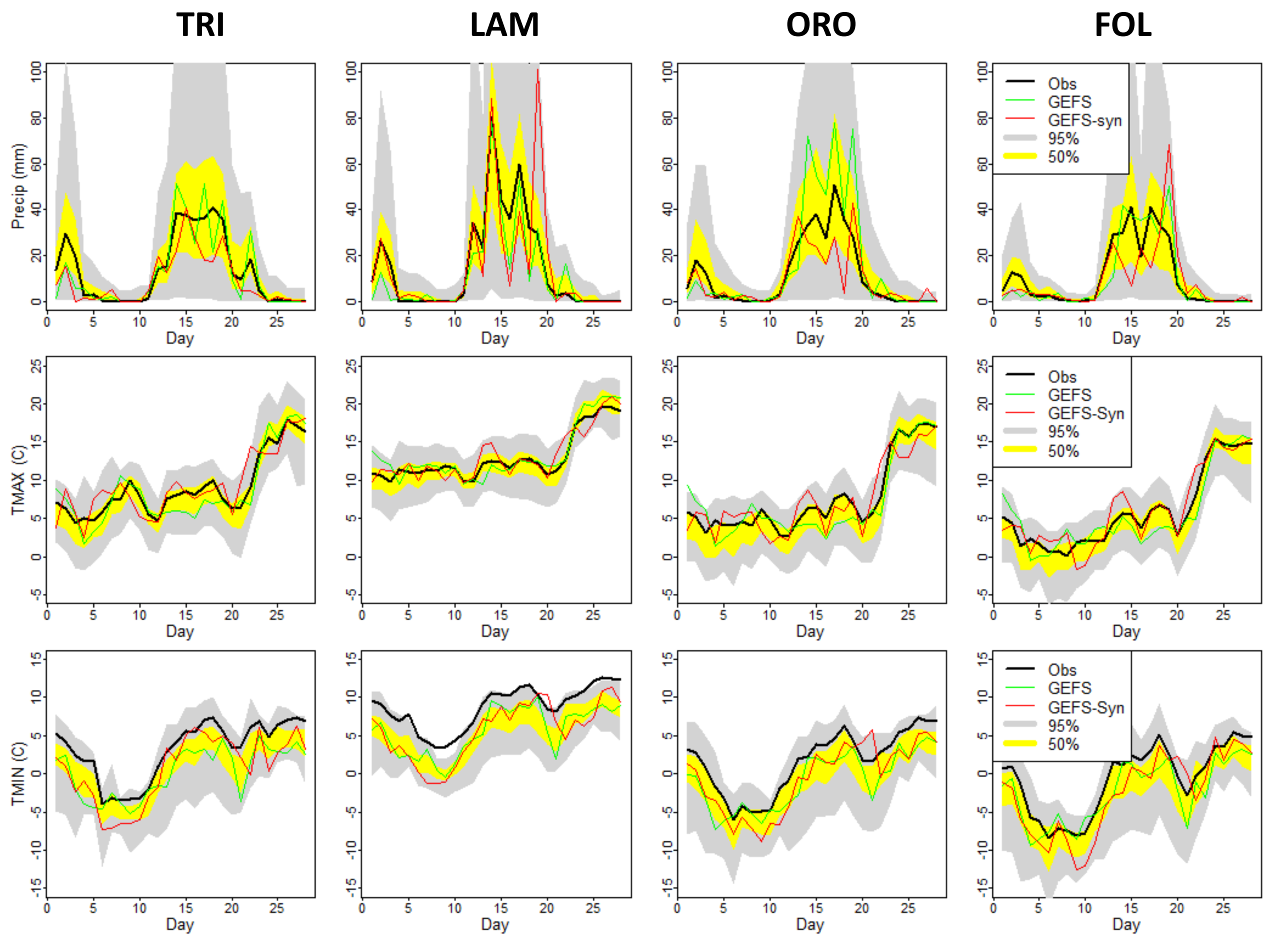

Figure 9: Synthetic forecast performance for February 1986 (extreme precipitation event mid-month) across 4 selected watersheds at a 5-day lead for variables of precipitation (PRECIP), maximum temperature (TMAX), and minimum temperature (TMIN). Observed values are indicated by solid black line, the GEFS ensemble mean forecast by solid green line, and the synthetic forecast (single random sample) by the solid red line. Grey (yellow) shading show 95th 95th (50th) percentile bounds across 1000 synthetic samples.

Figure 10 shows forecast validation metrics across sites for PRECIP, where reliability and skill are calculated only for non-zero precipitation days. Reliability for precipitation (top row) is generally above the 95% target for all sites, especially for the lower percentile bins, suggesting the synthetic forecasts are somewhat over-dispersed. Some slight dips in reliability occur across sites and lead times near the middle percentile bins, but the only noticeable excursions below the 95% line occur at LAM (1 and 5-day lead) and FOL (1-day lead). PRECIP forecast skill (2nd row) is summarized using a mean absolute error climatological skill score, which is similar to the metric used for streamflow but with absolute instead of squared errors (Wilks, 2011). The synthetic bounds for precipitation skill closely match that of the actual forecast, with only one exception of slightly over-estimated skill at shorter lead times for the ORO site. Uncertainty bounds for the skill metric are much tighter than those for the streamflow case study (Figure 7), largely because the sample size of observations is much larger for the meteorological case study (all data within the cold season, rather than just one month). Finally, we assess the ability of synthetic precipitation forecasts to replicate false positive and false negative PRECIP rates in the last two rows of Figure 10. As noted in section 3.4, these occurrence-based attributes of precipitation are sampled along with the synthetic a_t values, so we are primarily validating the meteorological KNN sampling procedure. Both false positive (row 3) and false negative (row 4) behavior is maintained accurately.

Figure 11 shows reliability and skill for TMAX forecasts. Reliability is somewhat higher than the 95% target at a 5-day lead at all sites (i.e., over-dispersed), but is near or sometimes slightly below the target at a 1-day lead. This likely results from the more Gaussian behavior of the error distribution at a 1-day versus a 5-day lead (see supporting information Figure S9). The pattern of forecast skill across lead times is also closely matched by the synthetic TMAX forecasts, although with the latter biased slightly positive. Again, the uncertainty bounds are very tight in these skill figures due to a large sample size.

Overall, Figures 10 and 11 suggest that the synthetic meteorological forecasts accurately preserve many of the properties of the empirical hindcasts at multiple sites and lead times.

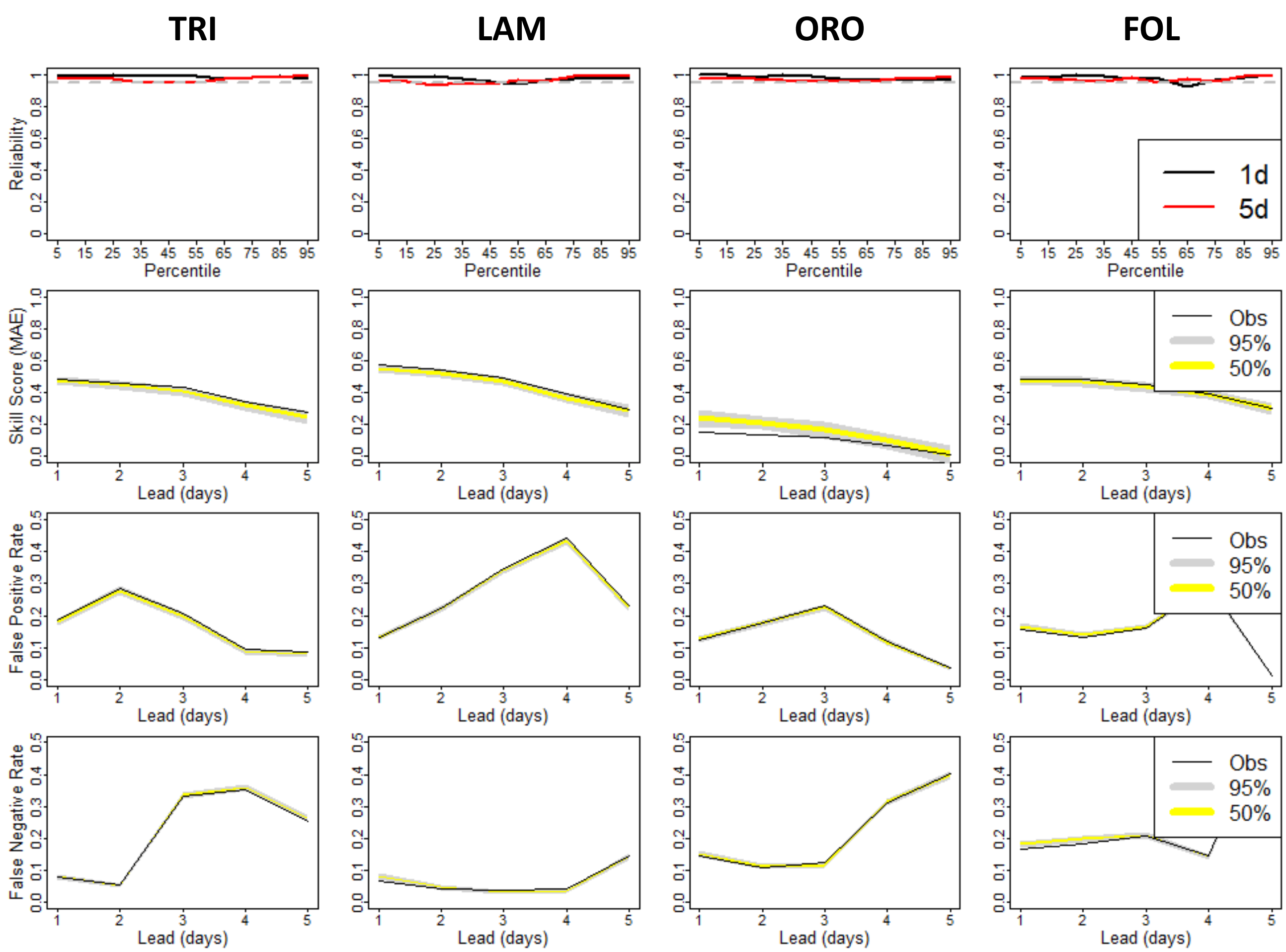

Figure 10: Precipitation (PRECIP) forecast metrics across 4 selected watershed grid cells. Top row – 95th percentile reliability plots; as in Figure 6, top row, but with 1-day (5-day) lead forecasts shown by black (red) solid line. 2nd row – Climatological skill score by mean absolute error (SS_{MAE}) across 5 forecast leads with the solid black line showing observed skill and gray (yellow) shading indicating 95th (50th) percentile bounds of 1000 synthetic samples. 3rd and 4th row: As in row 2, but showing false positive and false negative rates for rows 3 and 4 respectively.

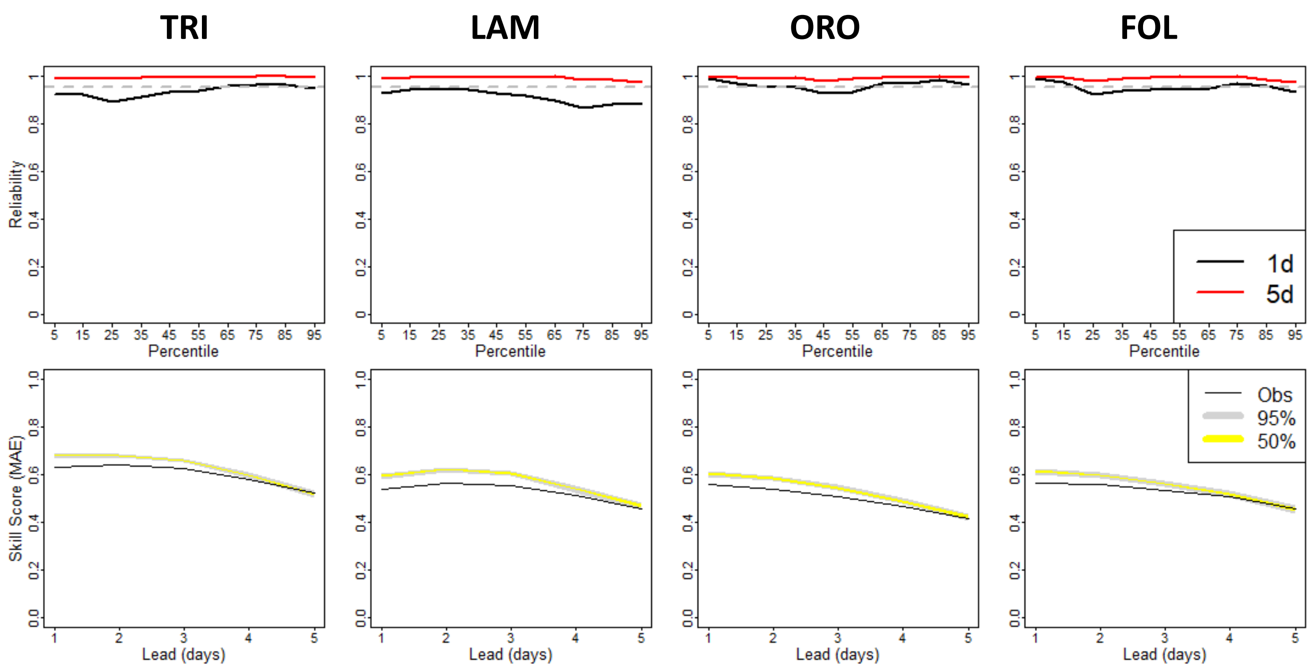

Figure 11: Cold season (ONDJFM) maximum temperature (TMAX) forecast metrics across 4 selected watershed grid cells. Top row – TMAX reliability as in Figure 9, top row. Bottom row – SS_{MAE} for TMAX as in Figure 9, row 2.

5. Discussion and Conclusion

This study contributes a generalized methodology to create synthetic forecasts from observed data that preserve empirical space-time and inter-variable relationships in forecast error. The methodology is adaptable to a wide variety of distributional forms and explicitly accounts for auto-correlation and heteroscedasticity in the forecast errors. We demonstrated short to medium range synthetic forecast generation in two case studies that highlighted the ability of the model to simulate accurately streamflow and meteorological forecasts across multiple lead times and locations from modern operational forecast systems. The two applications highlighted the model's potential for developing long records of streamflow forecasts via either the direct statistical approach or the conceptual hydrologic approach commonly used in designing, testing, and validating forecast informed water management policies (Lamontagne & Stedinger, 2018).

For the streamflow case study, we found that synthetic forecasts performed better (i.e., more accurately captured hindcast behavior) in the cold season months (ONDJFM) when observed streamflow was greater and more variable. The correspondence between the actual HEFS forecasts and observations deteriorated in the warm season (AMJJAS), causing a corresponding decrease in synthetic forecast performance. In general, our empirical copula and KNN sampling approach preserved key statistical relationships between the observations and the forecasts and between forecasts at different lead times. However, some challenges did remain; for instance when autocorrelation in forecast errors during large, infrequent flow events led to some unrealistically low synthetic forecasts during those events.

To support the conceptual hydrologic approach, we produced correlated meteorological forecasts of PRECIP, TMAX, and TMIN across locations and lead times. We found that our methodology readily adapted to the different distributional forms found in these variables and that synthetic output performed well against the actual forecasts. We also found that our sampling procedure, tailored specifically to capture the occurrence-based statistics of PRECIP, enabled an accurate representation of false positives and false negatives in the synthetic forecasts.

While the results of the two case studies were promising, there are avenues for model improvement that warrant additional discussion. In particular, modifications to the copula and sampling aspects of the methodology offer the potential for improved performance. We used an empirical copula, Schaake shuffle, and KNN sampling approach to explicitly preservecomplex correlation structures in the data. However, depending on the structure of the data, a parametric copula or multivariate kernel density estimate (Scott, 1992) of the empirical copula could also be employed, and this may allow for a richer characterization of forecast uncertainty across space and lead times. There are also other methods for bias correction, heteroscedastic modeling, and observational scaling that could be considered in future applications (see Schoups & Vrugt, 2010).

The flexibility of the modeling framework directly addresses critiques that research efforts in water resources often suffer from a lack of generalizability across spatio-temporal scales (Brown et al, 2015; Lall, 2014). While the case studies presented here focused on synthetic streamflow and meteorological forecasts, the model could be extended to other applications that require space-time scalability. For instance, recent work has stressed the importance of stochastic watershed modeling (Steinschneider et al., 2015; Vogel 2017) for long-term hydrologic risk assessment (rather than short-term forecasting). The model presented in this work could readily be extended to help develop correlated stochastic watershed models for river basins across a region, allowing for better characterization of risk in complex, multi-basin water systems. Furthermore, the adaptability of the methodology makes it well suited to exploratory efforts in forecast informed design like high-dimensional input/indicator variable selection (IVS) (Herman et al., 2020; Giuliani et al., 2015; Fernando et al., 2009), in addition to validation and testing of more established designs associated with FIRO (Jasperse et al., 2017; Delaney et al., 2020).

Lastly, we note that the proposed approach to synthetic forecast generation is applicable anywhere there is sufficient overlap in hindcast and observational data to fit the model. This confers the advantage of forecast record extension in areas with long observational records, but limited hindcasts. In cases where the observational record is also limited, there is the possibility of producing synthetic forecasts from traces of stochastically generated weather (Baxevani & Lennartsson, 2015; Steinschneider et al., 2019), which could significantly expand the data available to calibrate and test forecast-informed policies. This effort is left for future work.

Data Availability Statement

The data used in this manuscript are in the following repository and cited in the references: HydroShare, https://www.hydroshare.org/resource/4382404b935f4fde99c7ff4ada264867.

Acknowledgements

This study was supported by the U.S. National Science Foundation grant EnvS-1803563.

Reference

Alemu, E. T., Palmer, R. N., Polebitski, A., & Meaker, B. (2011). Decision Support System for Optimizing Reservoir Operations Using Ensemble Streamflow Predictions. *Journal of Water Resources Planning and Management*, *137*(1), 72–82. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000088

Anghileri, D., Voisin, N., Castelletti, A., Pianosi, F., Nijssen, B., & Lettenmaier, D. P. (2016). Value of long-term streamflow forecasts to reservoir operations for water supply in snow-dominated river catchments. *Water Resources Research*, *52*, 4209–42252. https://doi.org/10.1002/2015WR017864

Barth, R., Meibom, P., & Weber, C. (2011). Simulation of short-term forecasts of wind and load for a stochastic scheduling model. *In Proceedings of the IEEE Power and Energy Society General Meeting*, *24–28 July 2011*. Detroit, MI, USA, 1–8.

Baxevani, A., & Lennartsson, J. (2015). A spatiotemporal precipitation generator based on a censored latent Gaussian field. *Water Resources Research*, *51*, 4338–4358. https://doi.org/10.1002/2015WR017200.A

Bottazzi, G., & Secchi, A. (2011). A new class of asymmetric exponential power densities with applications to economics and finance. *Industrial and Corporate Change*, *20*(4), 991–1030. https://doi.org/10.1093/icc/dtr036

Box, G. E. P., & Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis*. New York: Wiley.

Brodeur, Z. P. (2020). Data repository for: A generalized approach to generate synthetic short-to-medium range hydro-meteorological forecasts, HydroShare, http://www.hydroshare.org/resource/4382404b935f4fde99c7ff4ada264867

Brodeur, Z. P., Herman, J. D., & Steinschneider, S. (2020). Bootstrap aggregation and cross-validation methods to reduce overfitting in reservoir control policy search. Water Resources Research, 55, e2020WR027184. https://doi.org/ 10.1029/2020WR027184

Brown, C. M., Lund, J. R., Cai, X., Reed, P. M., Zagona, E. A., Ostfeld, A., et al. (2015). Scientific Framework for Sustainable Water Management. *Water Resources Research*, 6110–6124. https://doi.org/10.1002/2015WR017114.Received

Buckle, D. J. (1995). Bayesian inference for skewed stable distributions. *Journal of the American Statistical Association*, *90*(430), 605–613. https://doi.org/10.1063/1.3573610

Cerqueti, R., Giacalone, M., & Panarello, D. (2019). A Generalized Error Distribution Copula-based method for portfolios risk assessment. *Physica A: Statistical Mechanics and Its Applications*, *524*, 687–695. https://doi.org/10.1016/j.physa.2019.04.077

Chen, L., & Guo, S. (2019). *Copulas and its application in hydrology and water resources*. Singapore: Springer.

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B., & Wilby, R. (2004). The Schaake shuffle: A method for reconstructing space-time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, *5*(1), 243–262. https://doi.org/10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2

Delaney, C. J., Hartman, R. K., Mendoza, J., Dettinger, M., Delle Monache, L., Jasperse, J., et al. (2020). Forecast Informed Reservoir Operations Using Ensemble Streamflow Predictions for a Multipurpose Reservoir in Northern California. *Water Resources Research*, *56*(9). https://doi.org/10.1029/2019WR026604

Demargne, J., Mullusky, M., Werner, K., Adams, T., Lindsey, S., Schwein, N., et al. (2009). Application of Forecast Verification Science to Operational River Forecasting in the U.S. National Weather Service. *Bulletin of the American Meteorological Society*, *June*, 778–784. https://doi.org/10.1175/2008BAMS26I9.I

Demargne, J., Wu, L., Regonda, S. K., Brown, J. D., Lee, H., He, M., et al. (2014). The science of NOAA's operational hydrologic ensemble forecast service. *Bulletin of the American Meteorological Society*, *95*(1), 79–98. https://doi.org/10.1175/BAMS-D-12-00081.1

De Mello, P.E., Lu, N., & Makarov, Y. (2011). An optimized autoregressive forecast error generator for wind and load uncertainty study. *Wind Energy, 14*, 967–976. https://doi.org/10.1002/we.460

Denaro, S., Anghileri, D., Giuliani, M., & Castelletti, A. (2017). Informing the operations of water reservoirs over multiple temporal scales by direct use of hydro-meteorological data. *Advances in Water Resources*, *103*, 51–63. https://doi.org/10.1016/j.advwatres.2017.02.012

Dettinger, M. D., Ralph, F. M., Das, T., Neiman, P. J., and D. R. Cayan, 2011: Atmospheric rivers, floods and the water resources of California. *Water*, *3*(4), 445–478, https://doi.org/10.3390/w3020445.

Eiras-Barca, J., Ramos, A. M., Pinto, J. G., Trigo, R. M., Liberato, M. L. R., & Miguez-Macho, G. (2018). The concurrence of atmospheric rivers and explosive cyclogenesis in the North Atlantic and North Pacific basins. *Earth System Dynamics*, *9*(1), 91–102. https://doi.org/10.5194/esd-9-91-2018

Ernst, M.D. (1998). A multivariate generalized Laplace distribution, *Computational Statistics*, *13*, 227–232. https://doi.org/10.1007/PL00022717

Giuliani, M., Zaniolo, M., Castelletti, A., Davoli, G., & Block, P. (2019). Detecting the State of the Climate System via Artificial Intelligence to Improve Seasonal Forecasts and Inform

Reservoir Operations. *Water Resources Research*, *55*(11), 9133–9147. https://doi.org/10.1029/2019WR025035

Giuliani, M., Pianosi, F., & Castelletti, A. (2015). Making the most of data: An information selection and assessment framework to improve water systems operations. *Water Resources Research, 51*, 9073–9093, https://doi.org/10.1002/2015WR017200.A.

Guan, B., Waliser, D. E., Ralph, F. M., Fetzer, E. J., & Neiman, P. J. (2016). Hydrometeorological characteristics of rain-on-snow events associated with atmospheric rivers. *Geophysical Research Letters*, 43(6), 2964–2973, https://doi.org/10.1002/2016GL067978

Fernández, C., & Steel, M. F. J. (1998). On bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, *93*(441), 359–371. https://doi.org/10.1080/01621459.1998.10474117

Hamill, T. M., Bates, G. T., Whitaker, J. S., Murray, D. R., Fiorino, M., Galarneau, T. J., et al. (2013). NOAA's second-generation global medium-range ensemble reforecast dataset. *Bulletin of the American Meteorological Society*, *94*(10), 1553–1565. https://doi.org/10.1175/BAMS-D-12-00014.1

Hanak, E., Lund, J., Dinar, A., Gray, B., Howitt, R., Mount, J., et al. (2011). *Managing California's Water*. Retrieved from http://www.ppic.org/content/pubs/report/R_211EHR.pdf

Hartmann, D. L. (2016) *Global Physical Climatology*, *2nd Ed.* Waltham, MA: Elsevier.

Heath, D. C., & Jackson, P. L. (1994). Modeling the Evolution of Demand Forecasts With Application to Safety Stock Analysis In Production/Distribution Systems. *IIE Transactions (Institute of Industrial Engineers)*, *26*(3), 17–30. https://doi.org/10.1080/07408179408966604

Hecht, C. W. & Cordeira, J. M. (2017). Characterizing the influence of atmospheric river orientation and intensity on precipitation distributions over North Coastal California. *Geophysical Research Letters*, *44*, 9048–9058, https://doi.org/10.1002/2017GL074179.

Herman, J. D., Quinn, J. D., Steinschneider, S., Giuliani, M., & Fletcher, S. (2020). Climate Adaptation as a Control Problem: Review and Perspectives on Dynamic Water Resources Planning Under Uncertainty. *Water Resources Research*, *56*(2). https://doi.org/10.1029/2019WR025502

Hodge, B.M., Lew, D., Milligan, M., Holttinen, H., Sillanpää, S., Gómez-Lázaro, E., et al. (2012). *Wind Power Forecasting Error Distributions: An International Comparison. In Proceedings of the International Workshop on Large-Scale Integration of Wind Power into Power Systems as Well as on Transmission Networks for Offshore Wind Power Plants, 13–15 November 2012*. Lisbon, Portugal.

Holton, J. R., & Hakim, G. J. (2013). *An Introduction to Dynamic Meteorology, 5th ed.* Waltham, MA: Elsevier.

Jasperse, J., Ralph, F. M., Anderson, M., Brekke, L., Dillabough, M., Dettinger, M., et al. (2017) Preliminary viability assessment of Lake Mendocino forecast informed reservoir operations. *Forecast Informed Reservoir Operations Steering Committee*, *5 July 2017*, San Diego, CA, Scripps Institute Center for Western Weather and Water Extremes, https://cw3e.ucsd.edu/FIRO_docs/FIRO_PVA.pdf.

Johnson, R. A., & Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis, 3rd ed.* New Jersey, Prentice-Hall.

Lall, U., & Sharma, A. (1996). A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resources Research*, *32*(3), 679–693.

Lall, U. (2014). Debates—The future of hydrological sciences: A (common) path forward? One water. One world. Many climes. Many souls. *Water Resources Research*, *50*, 5335–5341. https://doi.org/10.1111/j.1752-1688.1969.tb04897.x

Lamontagne, J. R., & Stedinger, J. R. (2018). Generating Synthetic Streamflow Forecasts with Specified Precision. *Journal of Water Resources Planning and Management*, *144*(4), 04018007. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000915

Landgraf, A. J., & Lee, Y. (2015). Dimensionality Reduction for Binary Data through the Projection of Natural Parameters. *Technical Report 890*, Department of Statistics, Ohio State University. URL http://arxiv.org/abs/1510.06112

Landgraf, M. A. J. (2016). *logisticPCA: Binary dimensionality reduction.* R package version 0.2. https://cran.r-project.org/web/packages/logisticPCA

Lettenmaier, D. P. (1984). Synthetic streamflow forecast generation. *Journal of Hydraulic Engineering*, *110*(3), 277–289. https://doi.org/10.1061/(ASCE)0733-9429(1984)110:3(277)

Loucks, D. P., and Van-Beek, E. (2017) *Water Resources Systems Planning and Management: An Introduction to Methods, Models, and Applications.* Switzerland: Springer.

Nadarajah, S. (2005). A generalized normal distribution. *Journal of Applied Statistics*, *32*(7), 685–694. https://doi.org/10.1080/02664760500079464

Nelson, D. B. (1991). Conditional Heteroskedasticity in Asset Returns : A New Approach. *Econometrica*, *59*(2), 347–370.

Nicholson, W. B., Wilms, I., Bien, J., & Matteson, D. S. (2020). High dimensional forecasting via interpretable vector autoregression. *Journal of Machine Learning Research*, *21*, 1–52. https://www.jmlr.org/papers/volume21/19-777/19-777.pdf

Nicholson, W. B., Matteson, D. S., & Bien, J. (2019). *BigVAR: Dimension reduction methods for multivariate time series.* R Package version 1.0.6. https://cran.r-project.org/web/packages/BigVAR

NOAA/NWS California/Nevada River Forecast Center. (CA/NV RFC). Hydrologic Ensemble Forecast System (HEFS) streamflow forecast output, Folsom Reservoir, CA (FOLC1). Retrieved March 1 2020 from Brett Whitin, P.E., CA/NV RFC.

California Department of Water Resources. (CA/DWR). California Data Exchange Center (CDEC) Observed full natural flow streamflow output, Folsom Reservoir, CA (FOLC1). Retrieved March 1 2020. https://cdec.water.ca.gov.

NOAA Physical Sciences Laboratory. (NOAA PSL). NCEP Global Ensemble Forecast System (GEFS) data archive. Retrieved June 1 2020. https://psl.noaa.gov/forecasts/reforecast2/download.html

NOAA Physical Sciences Laboratory. (NOAA PSL). NOAA-CIRES-DOE 20th Century Reanalysis Version 3 (20CR V3). Retrieved June 1 2020. https://psl.noaa.gov/data/gridded/data.20thC_ReanV3.html

NWS Office of Hydrologic Development (NWS OHD). (2016). HEFS Overview and Getting Started, Version OHD-CORE-CHPS-4.4.a, Release May 2016.

Olauson, J., Bladh, J., Lönnberg, J., & Bergkvist, M. (2016). A new approach to obtain synthetic wind power forecasts for integration studies. *Energies*, *9*(10), 1–16. https://doi.org/10.3390/en9100800

Pelland, S., Galanis, G., & Kallos, G. (2013). Solar and photovoltaic forecasting through post-processing of the Global Environmental Multiscale numerical weather prediction model.

Progress in Photovoltaics: Research and Applications, *21*, 284–296. https://doi.org/10.1002/pip

Piani, C., & Haerter, J. O. (2012). Two dimensional bias correction of temperature and precipitation copulas in climate models. *Geophysical Research Letters*, *39*(20), 1–6. https://doi.org/10.1029/2012GL053839

Raso, L., Schwanenberg, D., van de Giesen, N. C., & van Overloop, P. J. (2014). Short-term optimal operation of water systems using ensemble forecasts. *Advances in Water Resources*, *71*, 200–208. https://doi.org/10.1016/j.advwatres.2014.06.009

Rayner, S., Lach, D., & Ingram, H. (2005). Weather forecasts are for wimps: Why water resource managers do not use climate forecasts. *Climatic Change*, *69*(2–3), 197–227. https://doi.org/10.1007/s10584-005-3148-z

Schoups, G., & Vrugt, J. A. (2010). A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. *Water Resources Research*, *46*(10), 1–17. https://doi.org/10.1029/2009WR008933

Scott, D. W. (1992). *Multivariate Density Estimation.* New York: Wiley.

Slivinski, L. C., Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Giese, B. S., McColl, C., et al. (2019). Towards a more reliable historical reanalysis: Improvements for version 3 of the Twentieth Century Reanalysis system. *Quarterly Journal of the Royal Meteorological Society*, *145*(724), 2876–2908. https://doi.org/10.1002/qj.3598

So, M. K. P., Chen, C. W. S., Lee, J. Y., & Chang, Y. P. (2008). An empirical evaluation of fat-tailed distributions in modeling financial time series. *Mathematics and Computers in Simulation*, *77*(1), 96–108. https://doi.org/10.1016/j.matcom.2007.02.008

Steinschneider, S., Wi, S., & Brown, C. (2015), The integrated effects of climate and hydrologic uncertainty on future flood risk assessments. *Hydrologic Processes*, *29*, 2823– 2839. https://doi.org/10.1002/hyp.10409.

Steinschneider, S., Ray, P., Rahat, S. H., & Kucharski, J. (2019). A Weather-Regime-Based Stochastic Weather Generator for Climate Vulnerability Assessments of Water Systems in the Western United States. *Water Resources Research*, *55*(8), 6923–6945. https://doi.org/10.1029/2018WR024446

Subbotin, M. T. (1923). On the law of frequency of error. *Mat. Sb.*, *31*(2), 296–301.

Sun, M., Feng, C., & Zhang, J. (2020). Probabilistic solar power forecasting based on weather scenario generation. *Applied Energy*, *266*(December 2019), 114823. https://doi.org/10.1016/j.apenergy.2020.114823

Teegavarapu, R.S.V, Salas, J.D, & Stedinger, J.R. (Eds.) (2019). *Statistical Analysis of Hydrologic Variables: Methods and applications.* Reston, VA: American Society of Civil Engineers.

Turner, S. W. D., Bennett, J. C., Robertson, D. E., & Galelli, S. (2017). Complex relationship between seasonal streamflow forecast skill and value in reservoir operations. *Hydrology and Earth System Sciences*, *21*(9), 4841–4859. https://doi.org/10.5194/hess-21-4841-2017

Valeriano, O. C. S., Koike, T., Yang, K., Graf, T., Li, X., Wang, L., & Han, X. (2010). Decision support for dam release during floods using a distributed biosphere hydrological model driven by quantitative precipitation forecasts. *Water Resources Research*, *46*(10). https://doi.org/10.1029/2010WR009502

Rao, C. R., & Vinod, H. D. (Eds.). (2019). *Handbook of Statistics, Volume 41, Conceptual econometrics using R*. Waltham, MA: Elsevier.

Vogel, R.M. (2017). Stochastic watershed models for hydrologic risk management. *Water Security*, *1*, 28-35. https://doi.org/10.1016/j.wasec.2017.06.001

Wei, W. W. S. (2019) *Multivariate Time Series Analysis and its Applications.* New York: Wiley.

Wilks, D. S., (2011). *Statistical Methods in the Atmospheric Sciences, 3rd ed.* Waltham, MA: Elsevier.

Wilks, D. S. (2015). Multivariate ensemble Model Output Statistics using empirical copulas. *Quarterly Journal of the Royal Meteorological Society*, *141*(688), 945–952. https://doi.org/10.1002/qj.2414

Wurtz, D., Setz, T., Chalabi, Y., Boudt, C., Chausse, P., & Miklovac, M. (2020). *fGarch: Rmetrics – Autoregressive Conditional Heteroskedastic Modeling.* R package version 3042.83.2. https://cran.r-project.org/web/packages/fGarch

You, J. Y., & Cai, X. (2008). Determining forecast and decision horizons for reservoir operations under hedging policies. *Water Resources Research*, *44*(11), 1–14. https://doi.org/10.1029/2008WR006978

Yuan, X., Wood, E. F., & Ma, Z. (2015). A review on climate-model-based seasonal hydrologic forecasting: physical understanding and system development. *Wiley Interdisciplinary Reviews: Water*, *2*(5), 523–536. https://doi.org/10.1002/wat2.1088

Zhao, T., Zhao, J., Yang, D., & Wang, H. (2013). Generalized martingale model of the uncertainty evolution of streamflow forecasts. *Advances in Water Resources*, *57*, 41–51. https://doi.org/10.1016/j.advwatres.2013.03.008

Zimmerman, J. K. H., Carlisle, D. M., May, J. T., Klausmeyer, K. R., Grantham, T. E., Brown, L. R., & Howard, J. K. (2018). Patterns and magnitude of flow alteration in California, USA. *Freshwater Biology*, *63*(8), 859–873. https://doi.org/10.1111/fwb.13058

SUPPORTING INFORMATION FOR:
A generalized approach to generate synthetic short-to-medium range hydro-meteorological forecasts

Zachary P. Brodeur[1], Scott Steinschneider[2]
Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY

1. Graduate Research Assistant, 111 Wing Drive, Riley-Robb Hall, Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY, 14853. Email: zpb4@cornell.edu, Phone: 607-255-2155 (Corresponding Author).

2. Assistant Professor, 111 Wing Drive, Riley-Robb Hall, Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY, 14853. Email: ss3378@cornell.edu, Phone: 607-255-2155.

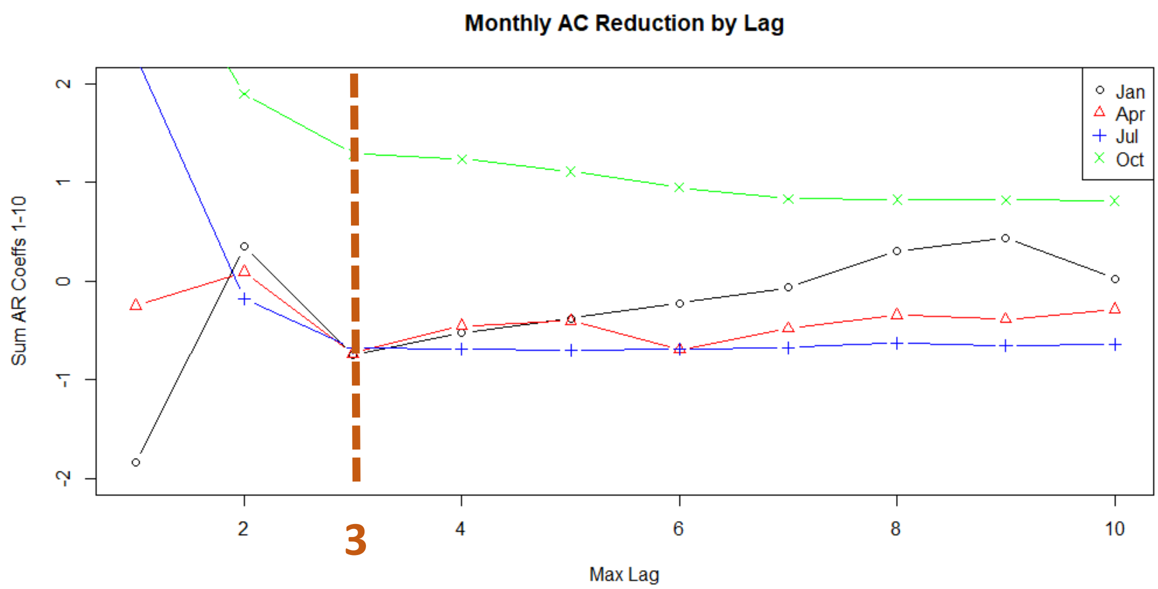

Figure S1: Summed auto-correlations coefficients (lags 1-10, all lead times) for Folsom reservoir forecast residuals after VAR decorrelation. The x-axis indicated the maximal lag-order in the BigVAR model and lines are shown for selected monthly subsets. Dark orange line demarcates maximal lag order chosen for this study.

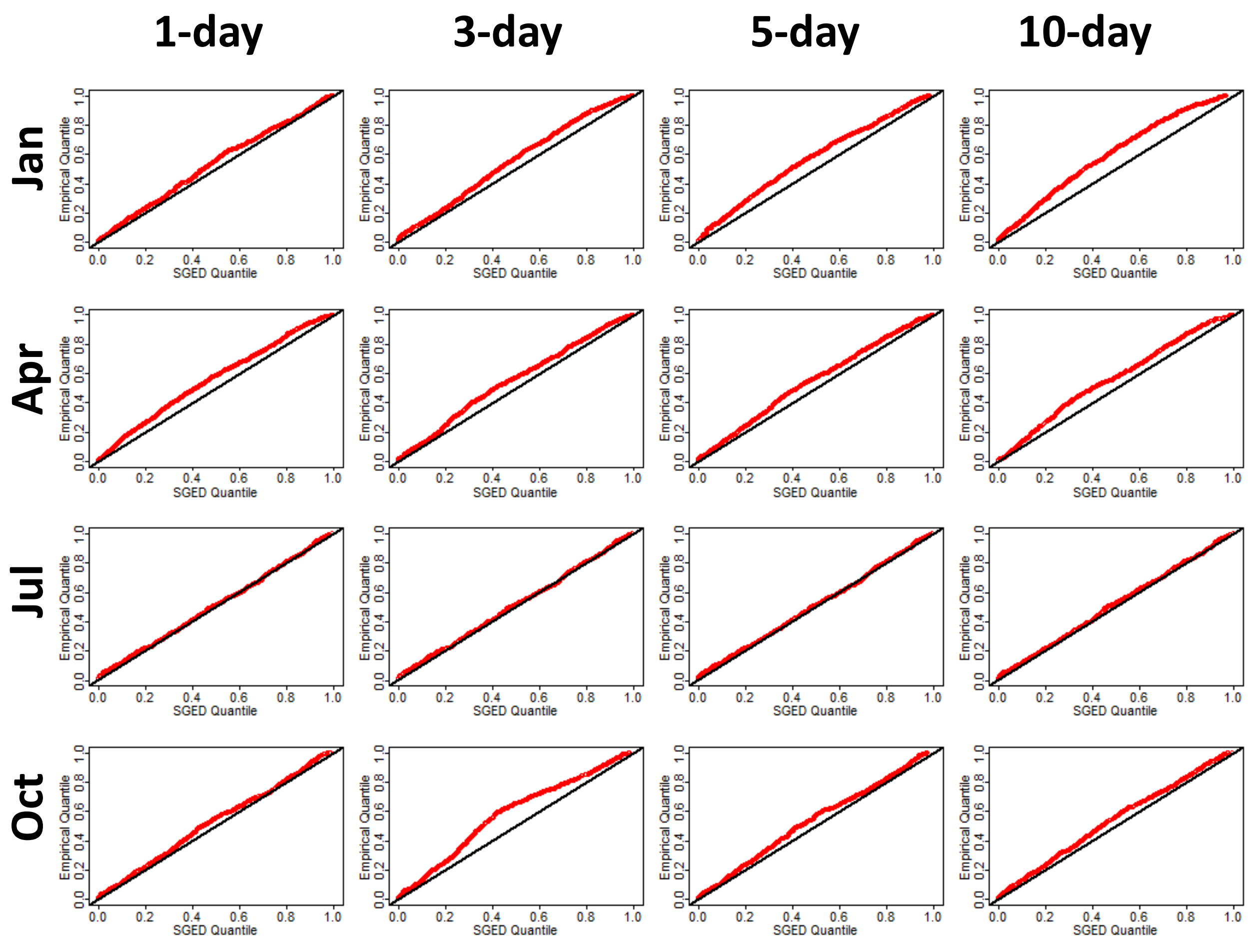

Figure S2: Q-Q plots across selected months (rows) and lead times (columns) for HEFS streamflow forecast transformed residuals (a_t). The black line is theoretical perfect correspondence between modeled and empirical quantiles (1:1) and red line shows the actual correspondence from the SGED model for the a_t residuals.

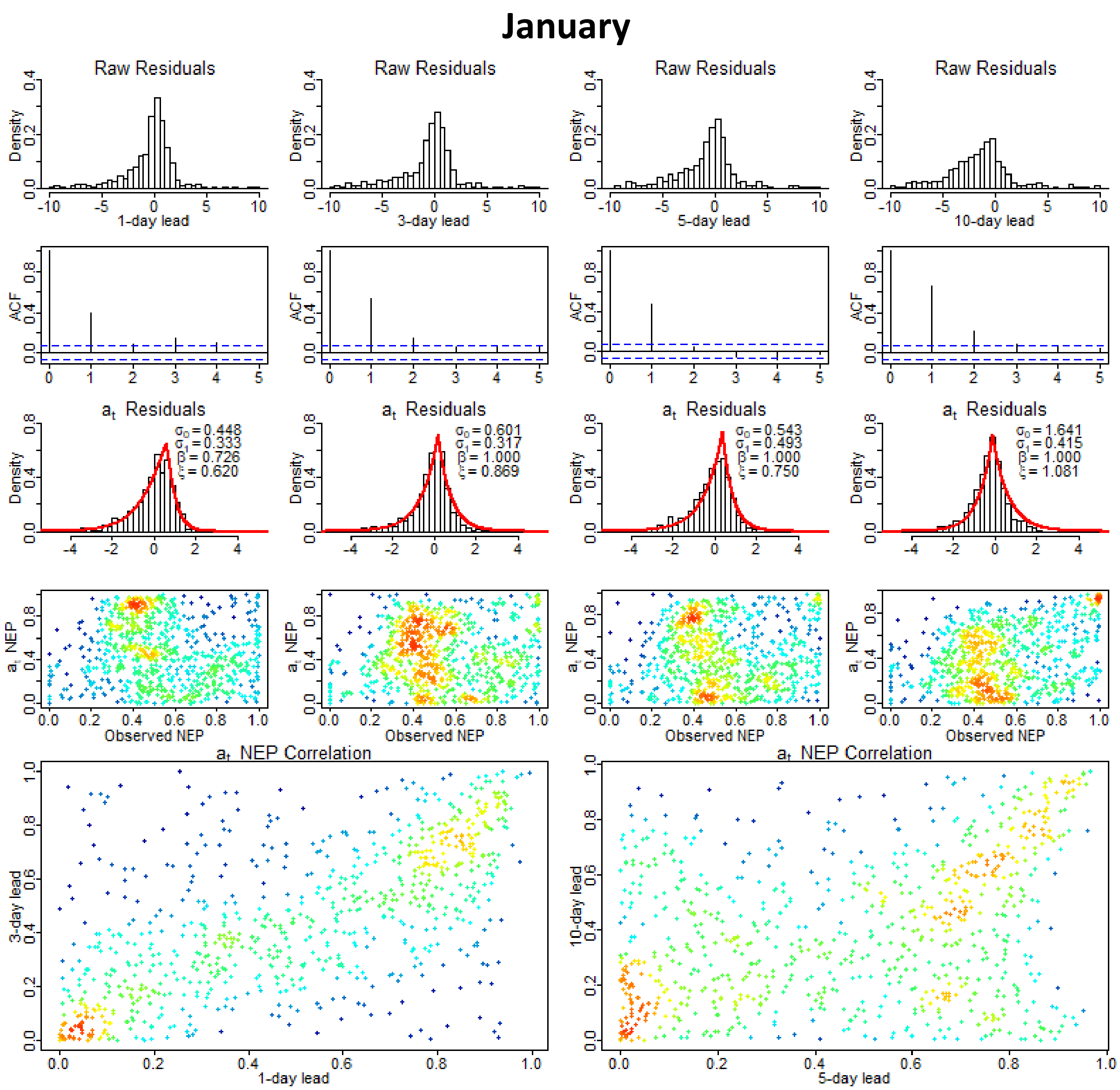

Figure S3: As in Figure 4, but only for month of January and across lead times of 1, 3, 5, and 10 days. Bottom row shows 1 to 3-day lead a_t NEP correlations in left panel and 5 to 10-day lead a_t NEP correlations in right panel.

April

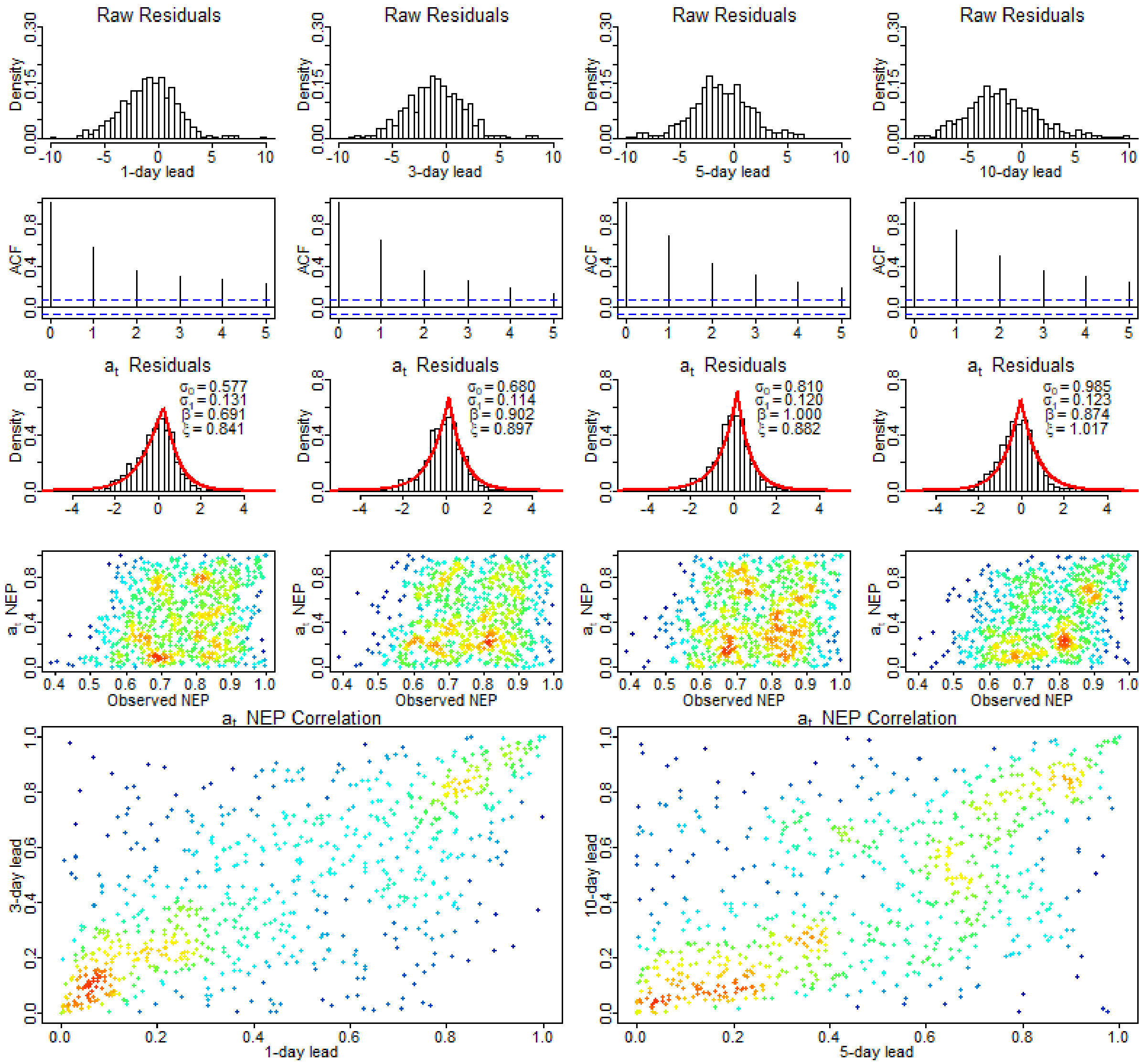

Figure S4: As in Figure 4, but only for month of April and across lead times of 1, 3, 5, and 10 days. Bottom row shows 1 to 3-day lead a_t NEP correlations in left panel and 5 to 10-day lead a_t NEP correlations in right panel.

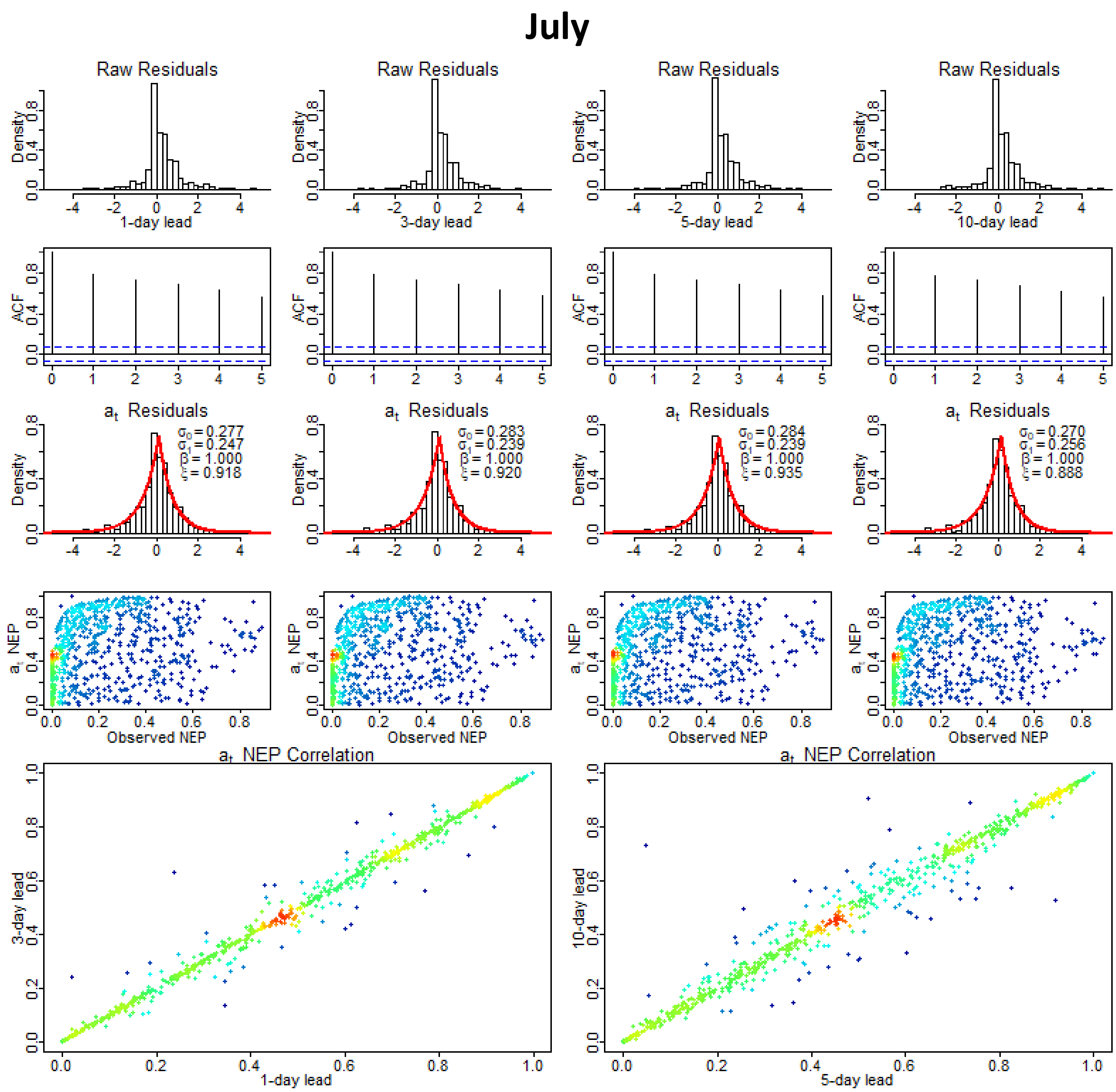

Figure S5: As in Figure 4, but only for month of July and across lead times of 1, 3, 5, and 10 days. Bottom row shows 1 to 3-day lead a_t NEP correlations in left panel and 5 to 10-day lead a_t NEP correlations in right panel.

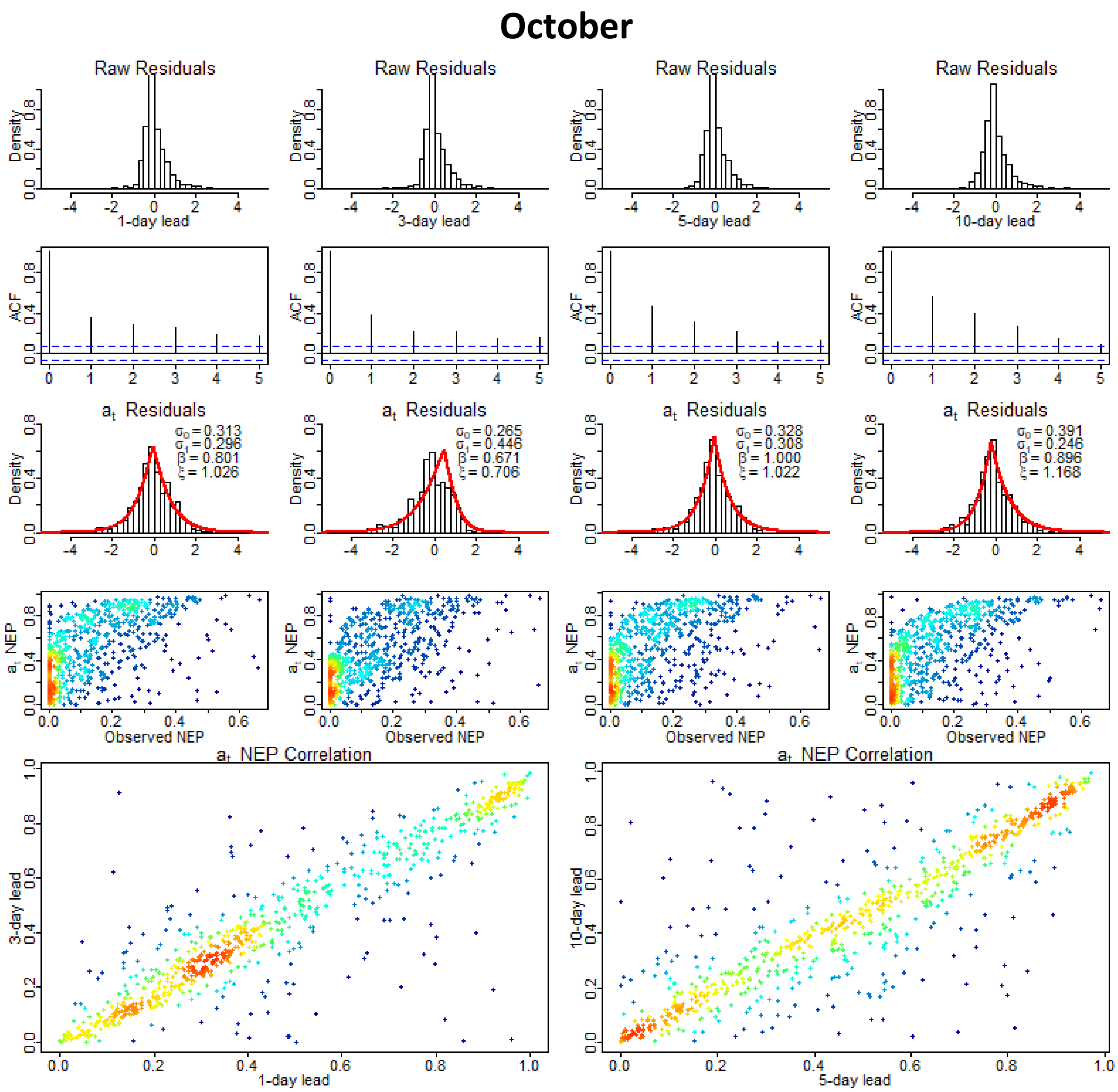

Figure S6: As in Figure 4, but only for month of October and across lead times of 1, 3, 5, and 10 days. Bottom row shows 1 to 3-day lead a_t NEP correlations in left panel and 5 to 10-day lead a_t NEP correlations in right panel.

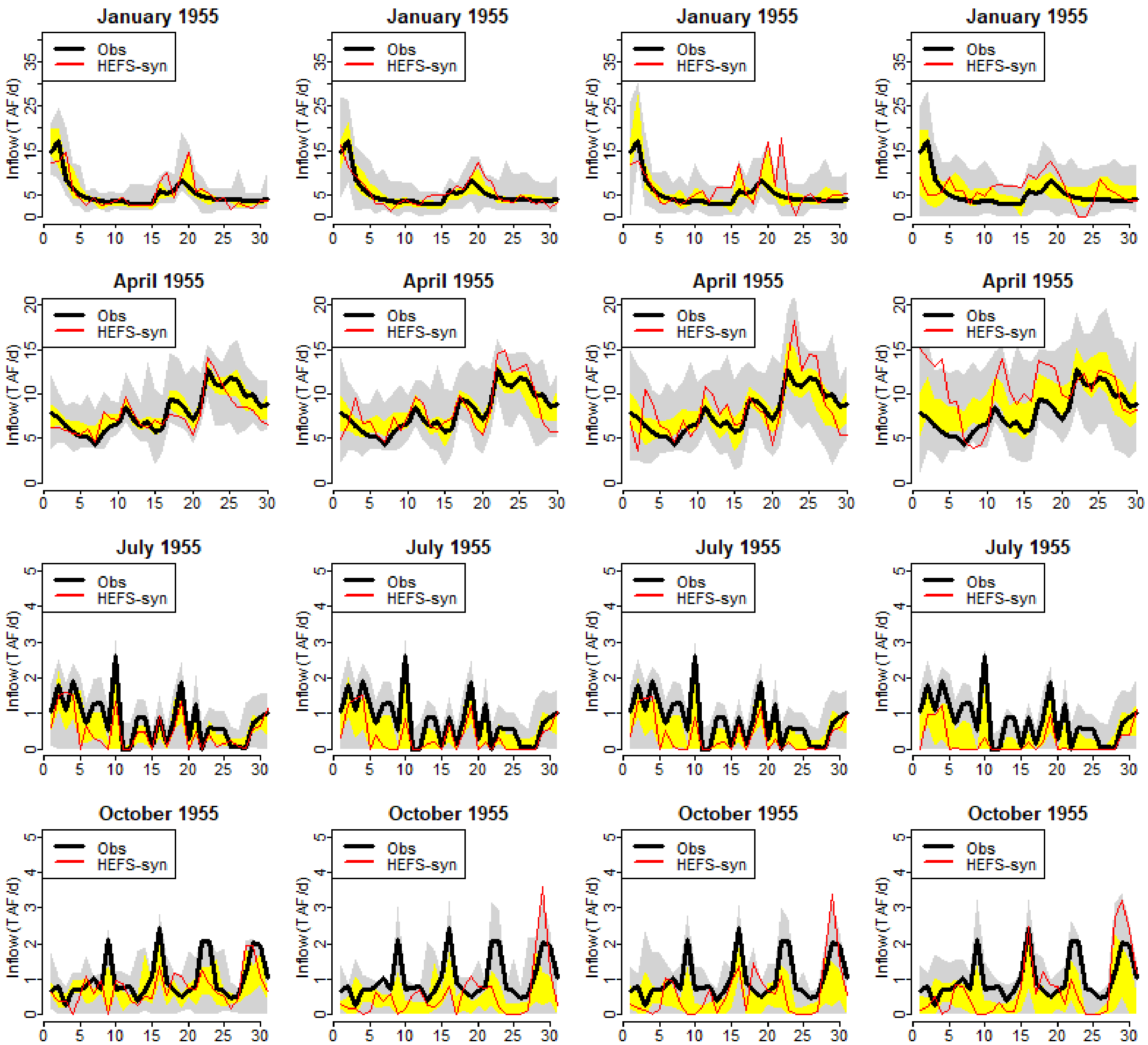

Figure S7: As in figure 6 but for 4 selected months in 1955 (i.e. synthetic period samples). Only observed full-natural-flow and synthetic data are shown since no actual HEFS hindcasts are available in this period.

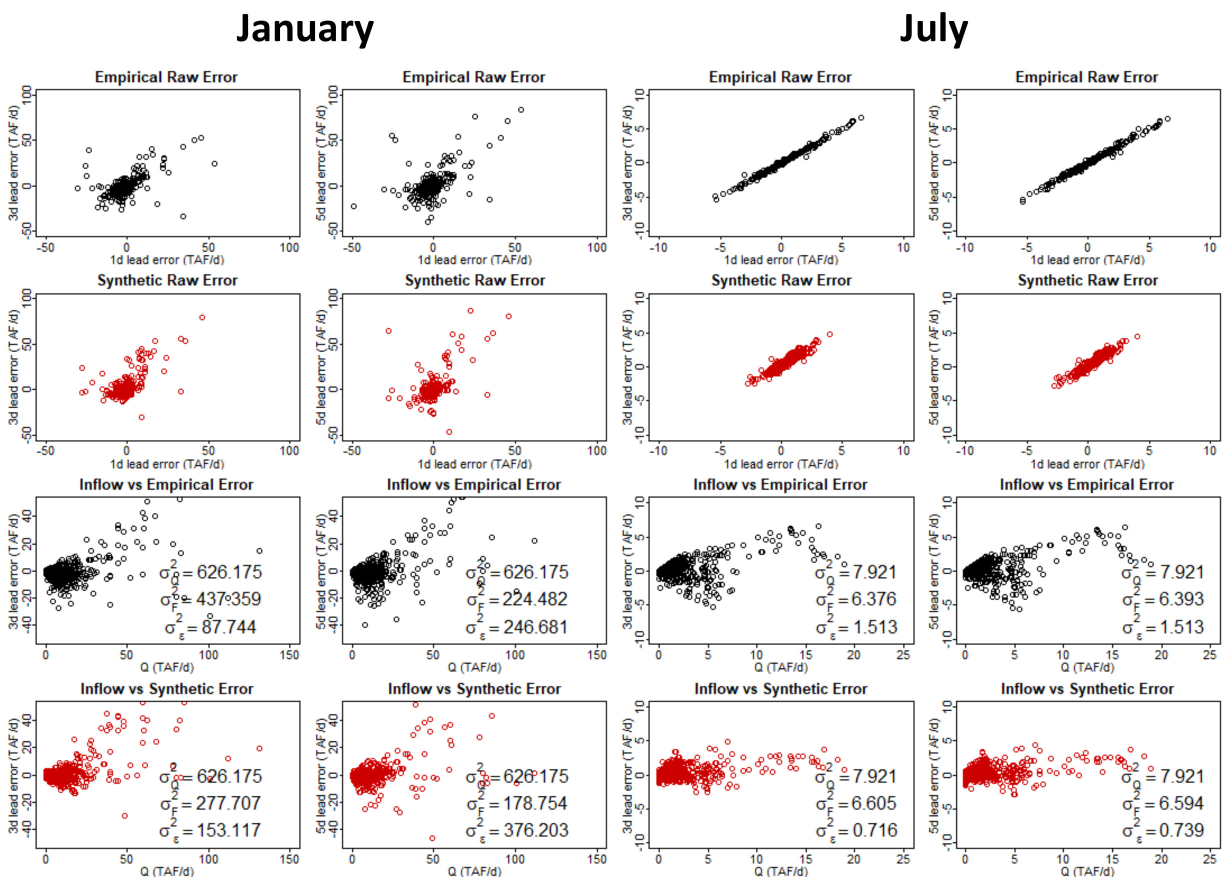

Figure S8: Raw error scatterplots for January (left two columns) and July (right two columns). Top row – Empirical raw error scatterplot (black) between 1 to 3-day forecast leads (1st and 3rd panel) and 1 to 5-day forecast leads (2nd and 4th panel). 2nd row – As for top row, but with synthetic raw error scatterplot (red). 3rd row – Observed flow versus empirical raw errors (black) at 3-day lead (1st and 3rd panel) and versus 5-day lead errors (2nd and 4th panel). Variance in observed flow ($\sigma_Q^{\,2}$), forecast ($\sigma_F^{\,2}$), and errors ($\sigma_\varepsilon^{\,2}$) are indicated top to bottom by text in the figure. Bottom row – As in 3rd row but for synthetic raw errors (red).

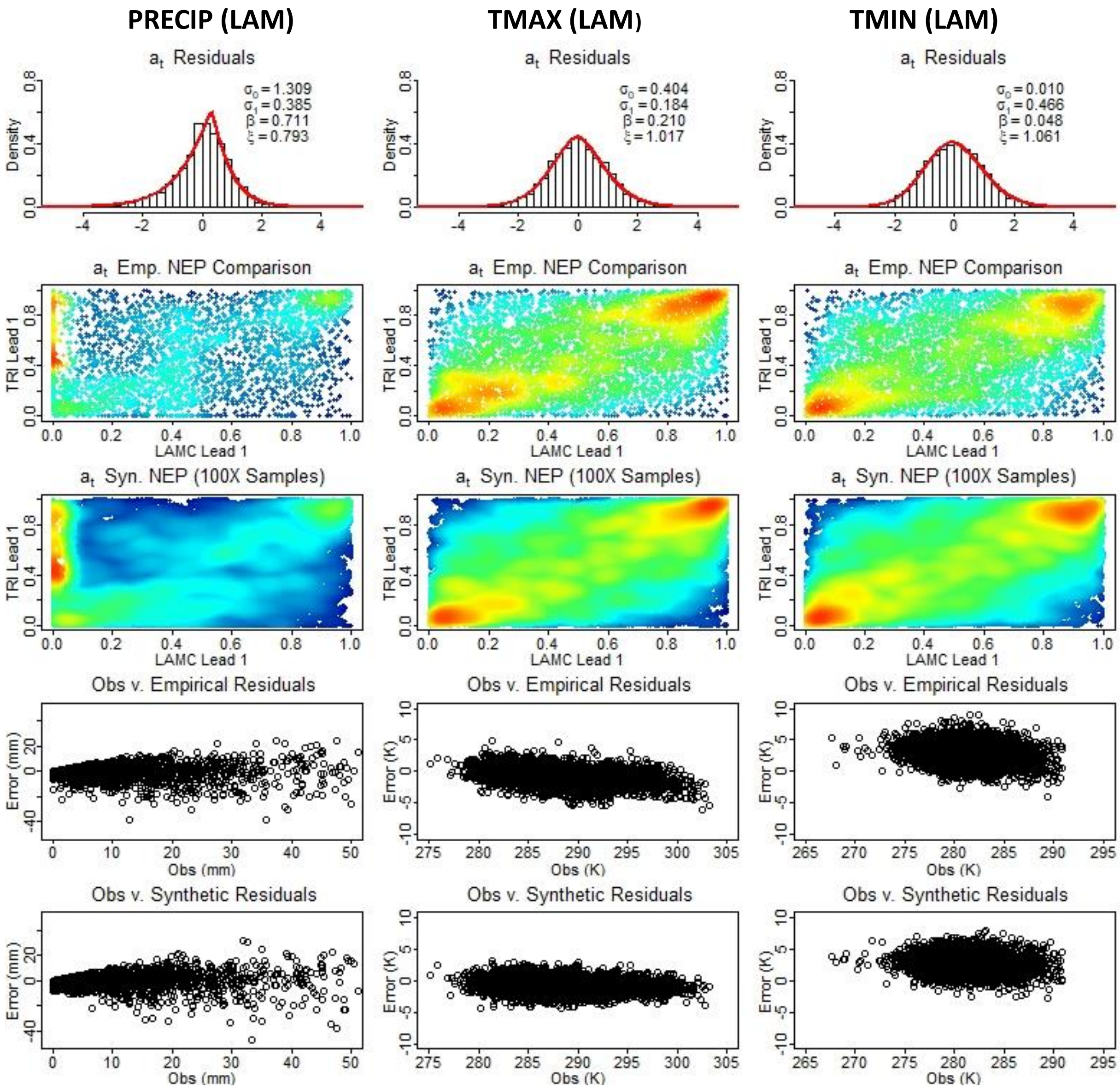

Figure S9: As in Figure 7 but for 1-day forecast lead and including TMIN column (rightmost)

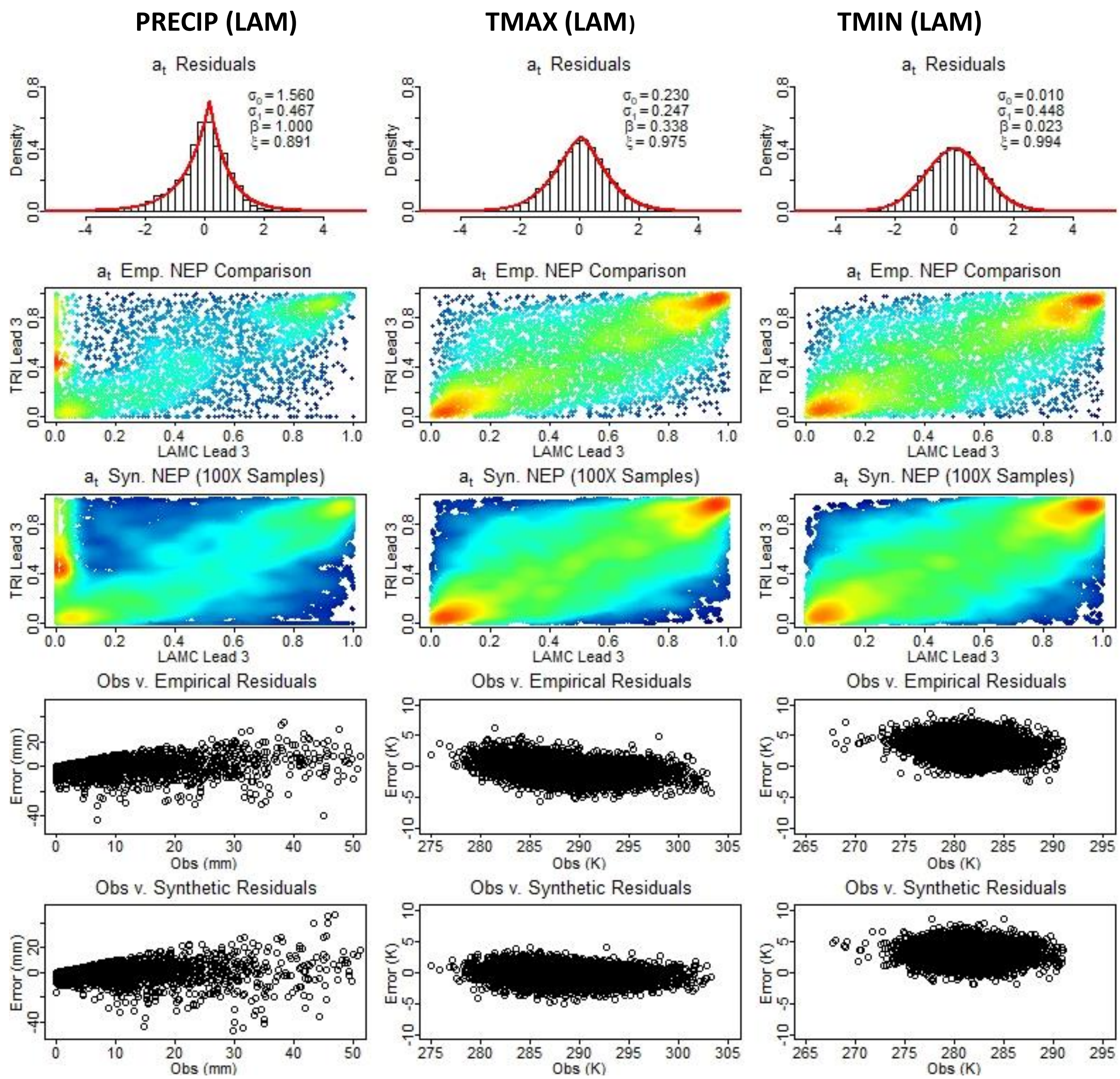

Figure S10: As in Figure 7 but for 3-day forecast lead and including TMIN column (rightmost)

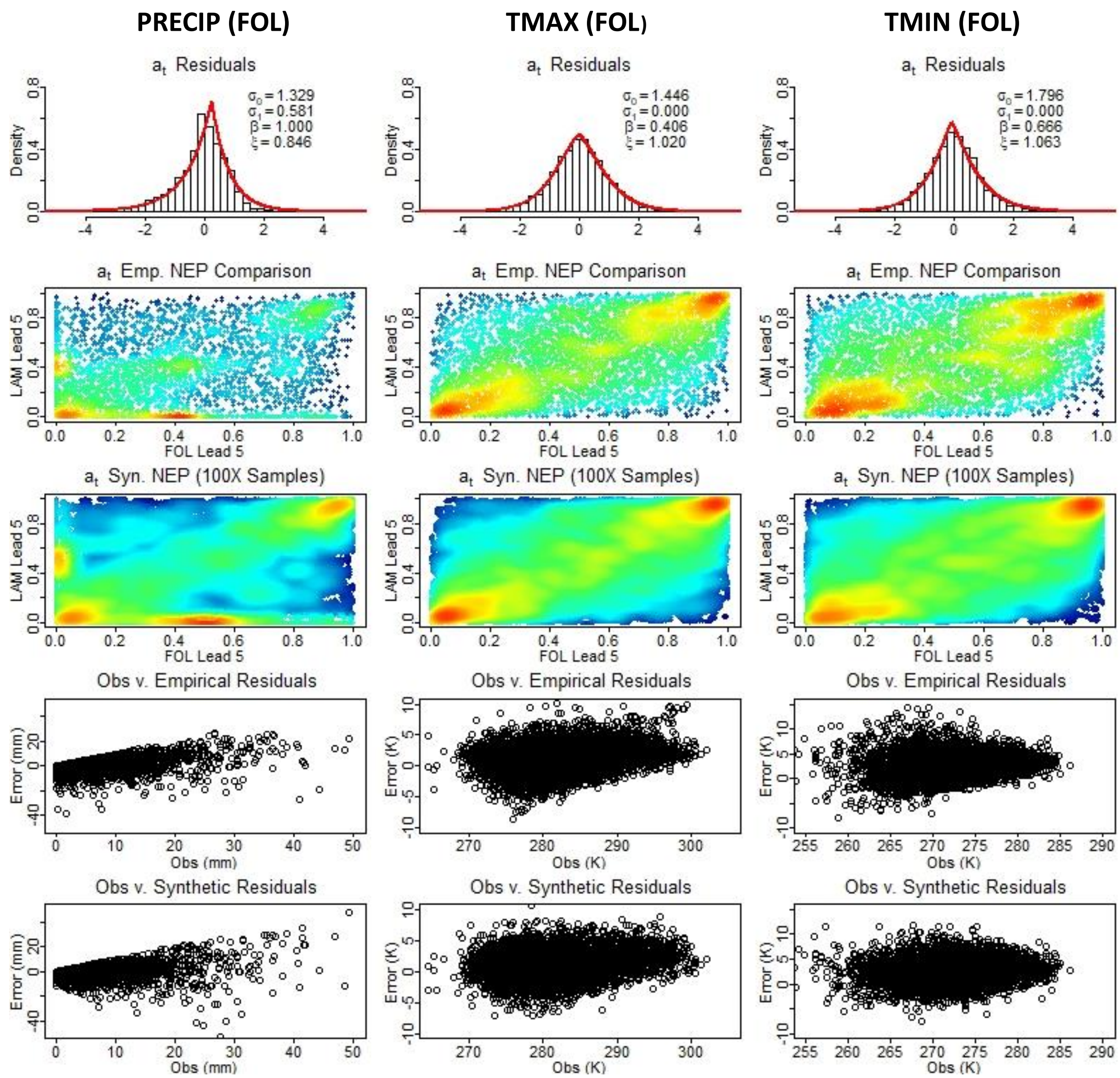

Figure S11: As in Figure 7 but with the Folsom Reservoir (FOL) grid cell compared against the Lake Mendocino (LAM) grid cell at a 5-day lead and including TMIN column (rightmost).

PRECIP (ORO) **TMAX (ORO)** **TMIN (ORO)**

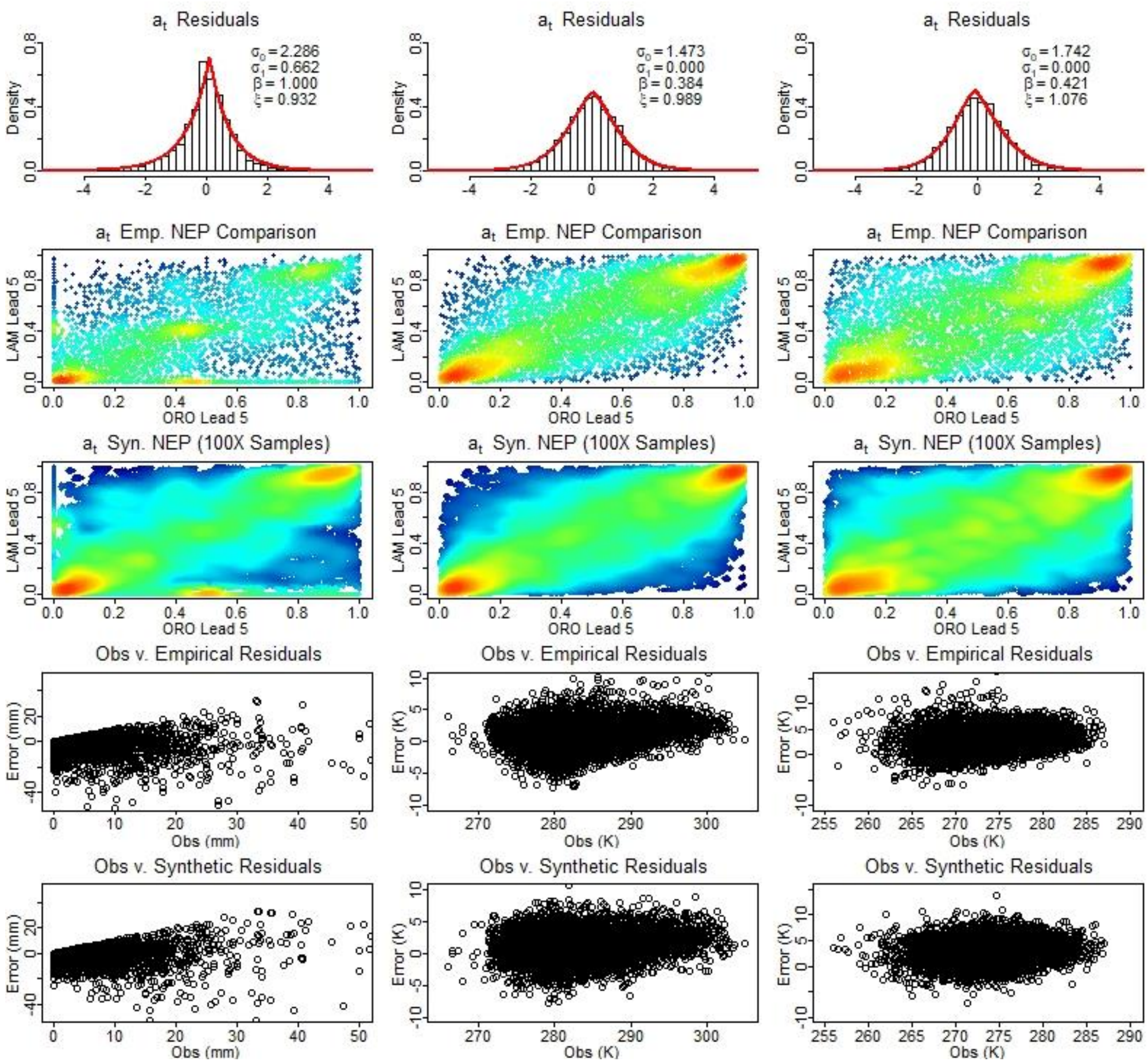

Figure S12: As in Figure 7 but with the Oroville Reservoir (ORO) grid cell compared against the Lake Mendocino (LAM) grid cell at a 5-day lead and including TMIN column (rightmost).

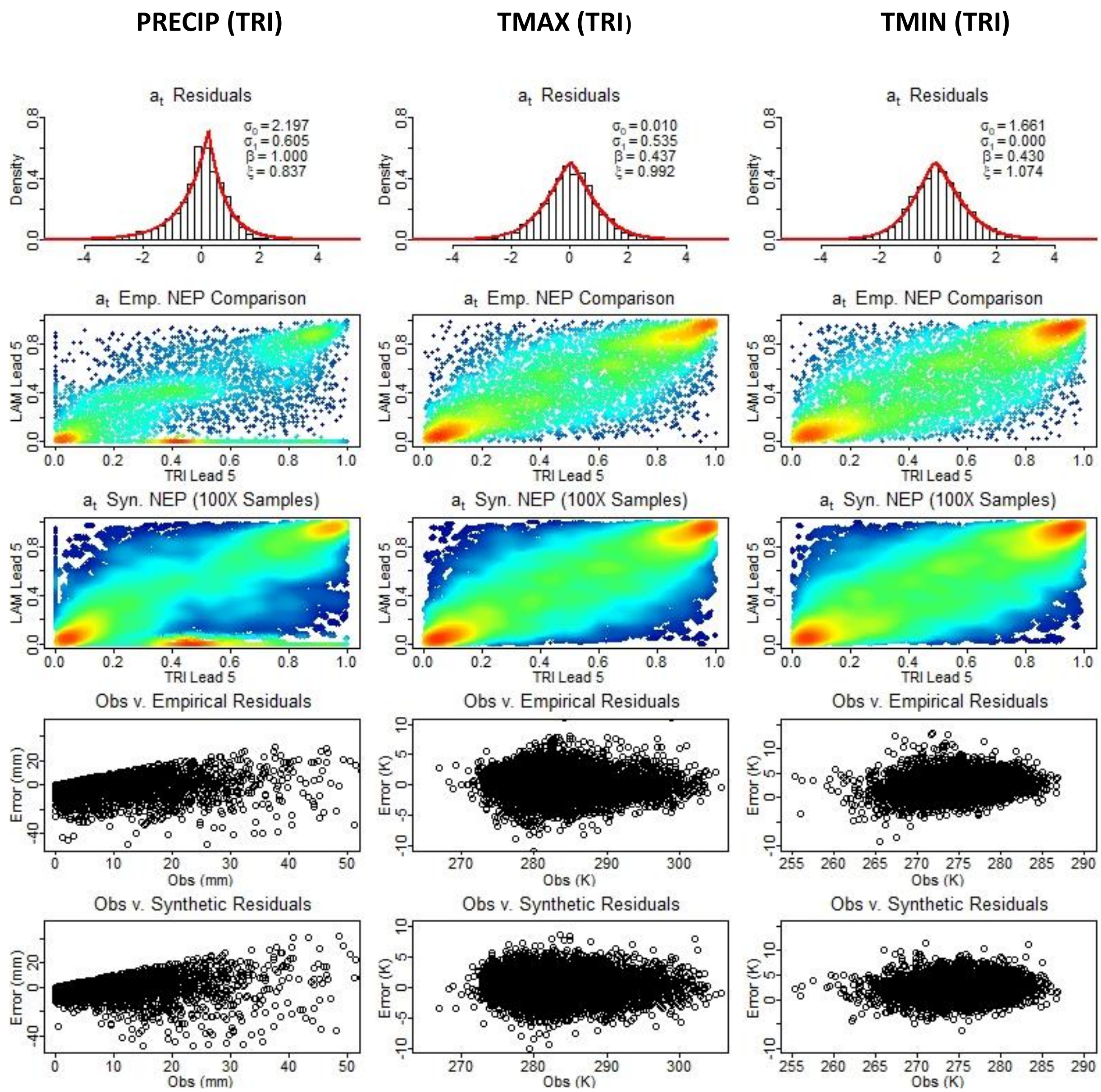

Figure S13: As in Figure 7 but with the Trinity Reservoir (TRI) grid cell compared against the Lake Mendocino (LAM) grid cell at a 5-day lead and including TMIN column (rightmost).

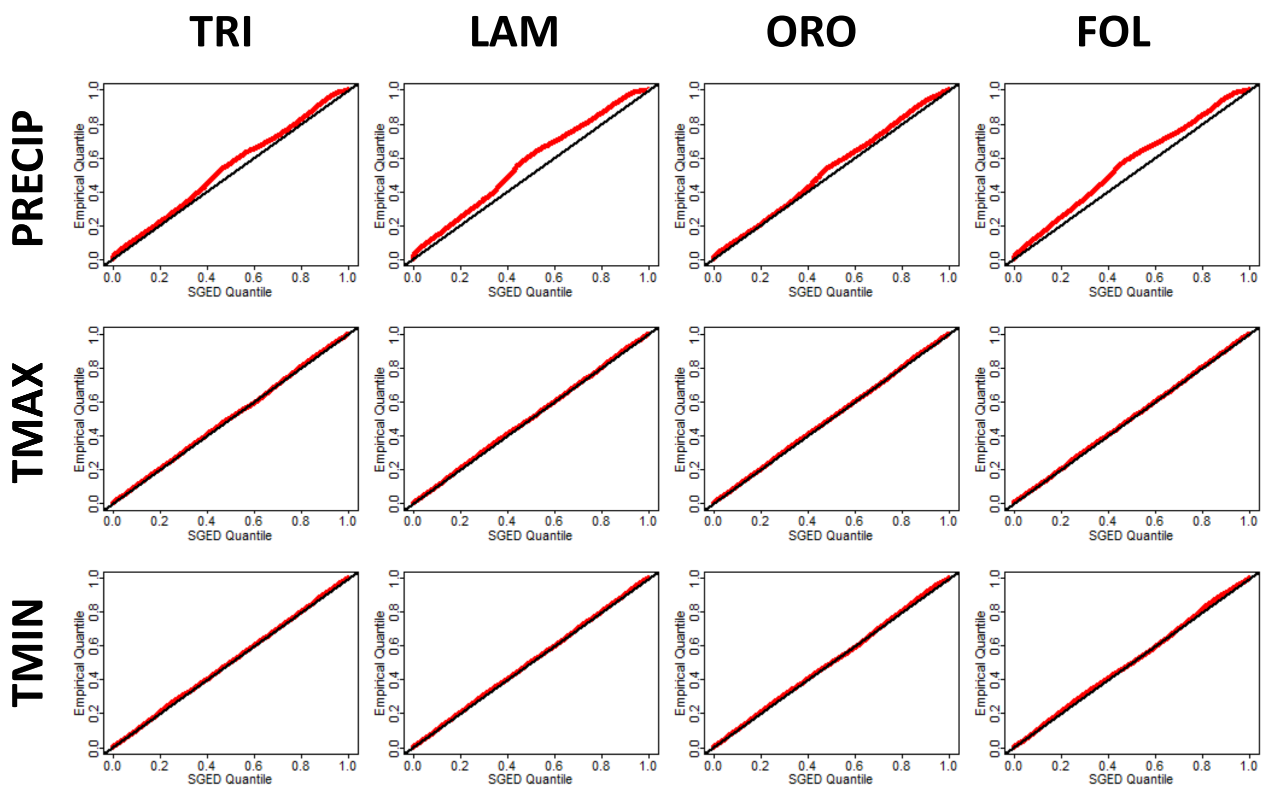

Figure S14: Q-Q plots across selected variables (rows) and basin grid cells (columns) for GEFS forecast transformed residuals (a_t) at a 5-day lead in cold-season (ONDJFM). The black line is theoretical perfect correspondence between modeled and empirical quantiles (1:1) and red line shows the actual correspondence from the SGED model for the a_t residuals.

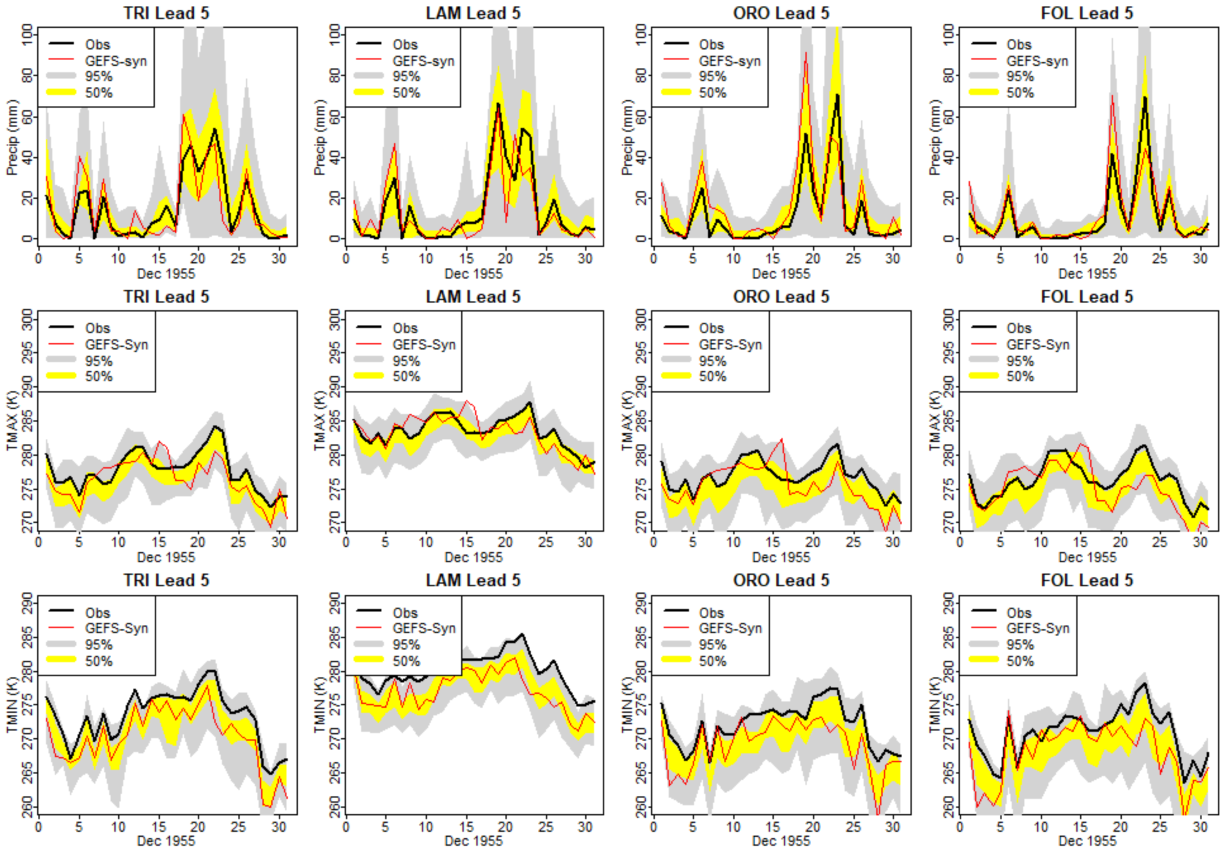

Figure S15: As in Figure 9 but for synthetic period (December 1955). Only observed (black) and synthetic forecast (red) values are shown since no GEFS hindcasts are available in this period.